TOWARDS DEVELOPMENT OF NO-REFERENCE OBJECTIVE MEASURES FOR PERCEPTUAL EVALUATION OF SINGING VOICE SEPARATION

A Dissertation
Presented to
The Academic Faculty

by

Udit Gupta

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
May 2017

TOWARDS DEVELOPMENT OF NO-REFERENCE OBJECTIVE MEASURES FOR PERCEPTUAL EVALUATION OF SINGING VOICE SEPARATION

Approved by:

Prof. Elliot Moore II, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Prof. Alexander Lerch, Co-Advisor
Center for Music Technology
Georgia Institute of Technology

Prof. Mark Clements
School of Electrical and Computer Engineering
Georgia Institute of Technology

Prof. Pamela Bhatti
School of Electrical and Computer Engineering
Georgia Institute of Technology

Prof. Timothy Hsu
Center for Music Technology
Georgia Institute of Technology

Prof. Justin Romberg
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: 6 January 2017

To my family,

their love and support helped me persevere.

ACKNOWLEDGEMENTS

This thesis is a consummation of years of hard work and dedication, and would not have been possible without the aid and support of my professors, my colleagues, and all the CSIP and ECE support staff.

I would like to specially thank my two advisers Prof. Elliot Moore and Prof. Alexander Lerch for their support, encouragement, and advice which made these last few years one of the best experiences of my life. I would also like to thank them for their mentoring and their steadfast belief in me for all this time.

I am grateful to the members of my dissertation committee Prof. Mark Clements, Prof. Pamela Bhatti, Prof. Justin Romberg, and Prof. Thimothy Hsu for their time and efforts on my behalf, as well as their invaluable insight and advice which helped my research become all the more meaningful. I would specially like to acknowledge Prof. Mark Clements and thank him for all the support and mentoring he provided me during my time here. Working with him on his projects allowed me to expand my expertise beyond the focus of my thesis and hopefully I am all the better researcher for it.

I would like to acknowledge all my colleagues and fellow students and I am grateful for all their help I received through out my time as a graduate student. All of you have been a large part of my life as a graduate student and helped make these years for me a very memorable experience.

A very special thanks to my parents and my sister as without their support and encouragement I would not be here.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Singing Voice Separation (SVS) uses audio source separation methods to isolate the vocal component from the background accompaniment in a song mix. A key challenge currently associated with evaluation of SVS is a lack of objective measures which correlate consistently with subjective evaluation. Additionally, the current state-of-the-art evaluation measures require the use of unmixed vocal and instrumental tracks which are often not available. The research presented in this thesis is an attempt to address these challenges by introducing two new objective measures for perceptually relevant evaluation of SVS.

The Vocal Isolation Score (VIS) is designed to assess the quality of isolation produced by various SVS algorithms when separating the vocals from the accompaniment. Similarly, the Vocal Intelligibility Preservation Score (VIPS) evaluates the preservation of intelligibility in the separated vocals. Other than an improvement upon the state-of-the-art, both VIS and VIPS have the additional advantage that they do not require references in the form of unmixed vocal or instrumental tracks to perform objective evaluation, unlike the currently popular objective measures used for evaluating audio source separation.

CHAPTER I

INTRODUCTION

Singing Voice Separation (SVS) gained prominence as a Music Information Retrieval (MIR) task due to its widespread use in other MIR tasks such as automatic lyrics recognition [66], singer identification [2, 17, 44], query by singing/humming [28], etc. It is also useful in the context of applications such as voice cancellation for karaoke, musical education, audio remixing, and rendering of stereophonic audio on multichannel systems. The goal for this task is to separate the vocals from the accompaniment in a song with mixed audio originally containing both. Various algorithms have been proposed which perform this task using diverse approaches [25, 29, 34, 50, 51, 52].

In Music Information Retrieval Evaluation eXchange (MIREX) competitions [10, 11, 8], the performance of various submissions for the SVS task was evaluated using (Normalized) Signal to Distortion Ratio (NSDR), Signal to Interference Ratio (SIR), and Signal to Artifacts Ratio (SAR) [62, 13]. However these measures (referred to collectively as BSS_Eval (Blind Source Separation Evaluation) measures in this work) do not correlate well with perceptual results obtained from human evaluation based on listening tests [21, 14, 5]. A brief overview of these objective measures for audio source separation, which constitute the current state-of-the-art, is presented in Chapter 2.

A key challenge currently associated with evaluation of SVS is a lack of objective measures which correlate consistently with subjective evaluation. Additionally, the current state-of-the-art evaluation measures, like BSS_Eval measures and PEASS (Perceptual Evaluation of Audio Source Separation) Measures, require the use of unmixed vocal and instrumental tracks which are often not available. The research

presented in this thesis is an attempt to address these challenges by introducing two new objective measures for evaluation of SVS without requiring the use of reference audio tracks containing the unmixed vocal or instrumental music. This thesis is an attempt at addressing these shortcomings of the current state-of-the-art measures by proposing new objective measures for perceptually motivated evaluation of separated vocals. To this end, a dataset comprising of results of perceptual evaluation of SVS from three listening experiments was prepared [21]. Chapter 3 discusses the details regarding the preparation, methodology, and analysis of the dataset and the results obtained are analyzed in Chapter 4.

Upon comparison between the subjective ratings obtained from the listening experiments and scores obtained from the objective measures, it was determined that both BSS_Eval and PEASS measures do not provide statistically significant correlations with the perceptual evaluation ratings (except for presence of artifacts) [21]. The discussion regarding the methodology used for the analysis and its results is presented in Chapter 5.

Two new objective measures Vocal Isolation Score (VIS) [19] and Vocal Intelligibility-Preservation Score (VIPS) [20] are proposed and discussed in Chapter 6 and Chapter 7 respectively. Both VIS and VIPS have been designed as "no-reference" measures, i.e. they do not need the unmixed vocal and instrumental tracks for the evaluation. The comparison of performance between these newly introduced measures and their analogues among the current state-of-the-art shows that both VIS and VIPS provide a remarkable improvement in performance over the existing measures in terms of consistency and accuracy.

The objective behind the research presented in this thesis is to address the gap in the community regarding perceptually relevant objective measures for evaluation of singing voice separation. This thesis demonstrates that the proposed measures perform better or as well as the established state-of-the-art in evaluating the quality

of separation of vocals from music. It also serves as an illustration of the fact that such evaluation is possible without the use of unmixed reference audio, there by bridging the gap between the use of academic audio datasets and real world applications.

CHAPTER II

BACKGROUND

This chapter provides a broad overview of the most common techniques used for extracting the vocal component from a mixed audio signal which contains overlapping vocal and instrumental music sources. In Section 2.1, some of the common impairments that may affect the quality of the extracted signals are discussed. An overview of the current state-of-the-art source separation and SVS evaluation techniques is provided in Section 2.2.

2.1 Common SVS Techniques and Related Impairments

Singing Voice Separation (SVS) is the process of recovering the original vocal source images from a signal containing a mix of vocal and instrumental music where the components overlap each other. It is analogous to the Blind Audio Source Separation (BASS) problem with an under-determined system due to music mostly being available as monophonic or stereophonic audio with multiple sources being mixed into together and the number of sources usually exceeding the number of channels (images) for the mixed audio. Unlike BASS, where sources are assumed to be independent, identically distributed (i.i.d.), and non-Gaussian, the techniques involved in SVS make informed decisions about the temporal and harmonic structures of the individual sources [41]. Additionally, in contrast to voice-only mixtures, musical sources can not be assumed to be mutually independent.

Music recordings can be of two types convolutive or instantaneous. The recordings that result from natural acoustic mixing (e.g. recording of a live concert) is convolutive in nature. On the other hand, most commercial recordings are produced in studios by synthetically mixing individual audio tracks. These recordings are said

to be instantaneous since most SVS systems consider added artificial reverberation as a part of the source signal. The SVS techniques presented in the following discussion are designed and tested using instantaneous mixtures due to the non-availability of the unmixed source signals in the case of convolutive mixtures.

Although there exists prior research for SVS based on exploiting the stereophonic nature of the spatial image (beam-forming) [55, 56], most of the state of the art approaches use a monophonic representation of the mixed audio signal for the purpose separation. The mixed audio signal is decomposed into some manner of time-frequency (TF) representation. Usually Short-time Fourier Transform (STFT) is used as the TF representation [25, 29, 34, 50, 51, 52, 68], however other approaches such as Lapped Orthogonal Transforms (LOTs) and Modified Discrete Cosine Transforms (MDCT) have also been used [49]. In typical SVS approaches, a TF mask is determined in order to isolate the partials corresponding to the fundamental frequency and the overtones for the vocals. Although this basic approach is common, the process used for determining the mask varies with different implementations [51, 29, 52]. The components of the audio signal that are passed through the mask constitute the separated vocal component, whereas the residual signal is the separated instrumental component. Section 2.1.1 discusses the more common methods for estimating time-frequency masks for vocal extraction and Section 2.1.2 presents some of the impairments associated with SVS.

2.1.1 Time-Frequency Mask Estimation Techniques for SVS

Various SVS algorithms use different techniques to determine the vocal component from the time-frequency (TF) representation of the audio. Some of the more common and successful techniques are presented in this section.

The first method involves exploiting the temporal and melodic continuity of the vocals. In this method, the algorithms try to determine which TF blocks correspond

to the vocal content. The most common technique is to find the TF blocks which lie at spectral peaks, and determine which of these TF blocks have the contributions of the vocals exceeding the contribution of the other sources by the way of exploiting harmonic relationships, and continuities in time and frequency that are characteristic for singing voices. Additional singing voice characteristics like vibrato and tremolo have also been used [24]. The mask can then be applied such that it allows the TF blocks containing the vocal signals to pass through and block everything else [52, 29, 24].

The second approach uses sparsity constraints for cost-function optimization in order to generate the TF mask. The SVS techniques using this idea suggest that there exist short-time repeating structures in the accompaniment portion of the music which is in contrast to the non-repeating structure of the vocals. These model the accompaniment as a low-rank subspace and use matrix factorization techniques such as Robust Principal Component Analysis (RPCA) [25], Similarity Matrix formulation [50, 51], Harmonic-Percussive Source Separation (HPSS) [34], or constrained Non-negative Matrix Factorization (NNMF) [46] to retrieve vocals from mixed audio signals.

There has also been some success reported with techniques for SVS which use a mix of both of the approaches [59, 15]. At this time, no comparative study could be found which discusses the merits of one approach over the other.

2.1.2 Common Impairments Associated with SVS

The common impairments observed in the audio extracted using SVS can be categorized into two types: ones which are audible due to the presence of other sources in the mix, and the ones that are introduced by the processing algorithm.

In the first category we have impairments due to the lack of "target preservation" that is a consequence of detection of false negatives in vocal detection. Instrumental

sources often may harmonically overlap with the vocal partials. This allows the instrumental content to leak through the TF mask as "source-interference."

The process used for singing voice separation and the methods used to apply the determined TF mask to extract the target signal also degrades the target audio. Partials corresponding to harmonics may be missed or spurious harmonics may be introduced in the TF mask, which leads to "target distortion." "Artifacts" such as musical-noise that degrade the audio are almost always introduced as a result of TF masking [30, 61].

2.2 SVS Evaluation - Current State of the Art

In order to consistently evaluate and compare the performance of different SVS algorithms, the use of a common scoring system is essential. There are many examples where subjective evaluation has been performed for comparing general audio source separation systems [36, 37, 68]. Many of these methods are geared towards evaluating source separation in speech-only mixtures, and hence do not transfer well to SVS evaluation where vocals are mixed with instrumental accompaniment. Additionally, listening tests are time-consuming, require carefully planning, and are usually restricted to a relatively small subset of audio files. To address these issues some objective methods for performance evaluation have been explored.

Vincent et al. [62] have suggested objective measures based on the presence of target spatial distortion (Image to Spatial Distortion Ratio, ISR), interference (SIR), and artifacts (SAR) in the separated signals as compared to the clean source signals. The total distortion in the output signal compared to the source is measured by Signal to Distortion Ratio (SDR) [1]. These measures model the test signal (extracted vocal) as a linear mixture of the target signal (true vocal), interfering signals (true instrumental), and noise (processing artifacts). The contribution of each of these sources is estimated to get the values for SIR, SAR, and (N)SDR. These measures

are a part of the publicly available BSS_Eval Toolbox [62] and will be collectively referred to as BSS_Eval measures. A detailed discussion of these measures is provided in Section 2.2.1. SIR and SAR were used to evaluate the submissions in the MIREX 2014 and 2015 Singing Voice Separation tasks, along with the normalized version of SDR designated as NSDR [10].

Another set of measures have been developed as improved versions of the BSS_Eval measures [14]. These measures, referred to as Overall Perceptual Score(OPS), Target-related Perceptual Score (TPS), Interference-related Perceptual Score (IPS), and Artifacts-related Perceptual Score (APS) are designed by taking the human loudness perception models into account and have been reported to have higher correlation scores as compared to the BSS_Eval measures. OPS, TPS, IPS, and APS are available as parts of the PEASS Toolbox [14, 63] and will be referred to as PEASS measures collectively henceforth. Detailed implementation of these measures is discussed in Section 2.2.2

Both BSS_Eval measures and PEASS measures are "Full Reference" measures, i.e., they require the true vocal signal and the true instrumental signal in the form of pre-mixed audio. A detailed evaluation of performance of both BSS_Eval measures and PEASS measures as compared to subjective results from an SVS listening test are discussed in Chapter 5.

2.2.1 BSS_Eval Measures - Implementation and Performance

BSS_Eval measures were developed by Vincent et. al. [62] for the purpose of comparison of Blind Audio Source Separation (BASS) techniques when used with identical datasets. They propose a general mixing model with n sources $\{s_1(t), \cdots, s_n(t)\}$; which are recorded on m channels $\{x_1(t), \cdots, x_m(t)\}$. Each channel can be modeled as a mixture of sources with additive noise, as shown in Eqn. (2.1). Here a_{ij} is a time

varying mixing filter and $n_i(t)$ is additive noise.

$$x_i(t) = \sum_{j=1}^{n} \sum_{\tau=0}^{\infty} a_{ij}(\tau) s_j(t-\tau) + n_i(t) \tag{2.1}$$

The estimated signal $\hat{s}_j$ can be decomposed into different components, as shown in Eqn. (2.2), regardless of the source separation methodology used. Here s_{target} is a function of the true source signal s_j, whereas e_{interf}, e_{noise}, and e_{artif} are error terms due to interfering sources, noise, and source separation artifacts.

$$\hat{s}_j = s_{\text{target}} + e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}} \tag{2.2}$$

The decomposition terms in Eqn. (2.2) can be calculated as orthogonal projections. Let $\Pi\{y_1, \cdots, y_k\}$ denote the orthogonal projection operator on a subspace spanned by the vectors $y_1, \cdots, y_k$. Three such operators are defined as:

$$P_{s_j} := \Pi\{s_j\} \tag{2.3}$$

$$P_s := \Pi\{(s_{j'})_{1 \le j' \le n}\} \tag{2.4}$$

$$P_{s,n} := \Pi\{(s_{j'})_{1 \le j' \le n}, (n_i)_{1 \le i \le m}\} \tag{2.5}$$

Therefore, the decomposition terms can be calculated as:

$$s_{\text{target}} = P_{s_j} \hat{s}_j \tag{2.6}$$

$$e_{\text{interf}} = P_s \hat{s}_j - P_{s_j} \hat{s}_j \tag{2.7}$$

$$e_{\text{noise}} = P_{s,n} \hat{s}_j - P_s \hat{s}_j \tag{2.8}$$

$$e_{\text{artif}} = \hat{s}_j - P_{s,n} \hat{s}_j \tag{2.9}$$

Various ratios of different combinations of the decomposition terms are used as the measures for objective evaluation of blind source separation systems, as shown in Equations (2.10-2.12). These measures are defined as the "Source to Distortion Ratio" (SDR)

$$\text{SDR} := 10 \log_{10} \frac{\|s_{\text{target}}\|^2}{\|e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}\|^2}, \tag{2.10}$$

the "Source to Interference Ratio" (SIR)

$$\mathrm{SIR} := 10 \log_{10} \frac{\|s_{\mathrm{target}}\|^2}{\|e_{\mathrm{interf}}\|^2}, \tag{2.11}$$

and the "Source to Artifacts Ratio" (SAR)

$$\mathrm{SAR} := 10 \log_{10} \frac{\|s_{\mathrm{target}} + e_{\mathrm{interf}} + e_{\mathrm{noise}}\|^2}{\|e_{\mathrm{artif}}\|^2}. \tag{2.12}$$

In the context of SVS evaluation, the assumption that the mixing gain is time invariant is not always true. Therefore, in the case of instantaneous mixtures with time varying gains, the performance measures SDR, SIR, and SAR are calculated locally by windowing the signals. The final measure can be evaluated as the average of the local measures. In the MIREX competition SVS Task [10, 11], an additional measure "Normalized Signal to Distortion Ratio" (NSDR) is used, which is calculated by finding the "Source to Distortion Ratio" (SDR^*), with x_{mix} used in place of $\hat{s}_{\mathrm{voc}}$ and $\hat{s}_{\mathrm{ins}}$ and subtracting the original SDR from it as

$$\mathrm{NSDR} := \mathrm{SDR}^* - \mathrm{SDR} \tag{2.13}$$

The BSS_Eval measures discussed in this section have been criticized in literature for their inability to fit subjective ratings [14, 63, 6, 21, 5]. In various listening tests performed for subjective evaluation of source separation [14, 6, 21] (both voice only mixtures, and vocal with instrumental mixtures), these measures show very little correlation with the listening test results.

2.2.2 PEASS Measures - Implementation and Performance

The PEASS Measures have been designed to overcome the limitation of BSS_Eval measures with auditory phenomena such as loudness perception and spectral masking being taken into account. In this case, the distortion between the estimate signal $\hat{s}_{ij}(t)$ and the target signal $s_{ij}(t)$ is decomposed as

$$\hat{s}_{ij}(t) - s_{ij}(t) = e_{ij}^{\mathrm{target}}(t) + e_{ij}^{\mathrm{interf}}(t) + e_{ij}^{\mathrm{artif}}(t), \tag{2.14}$$

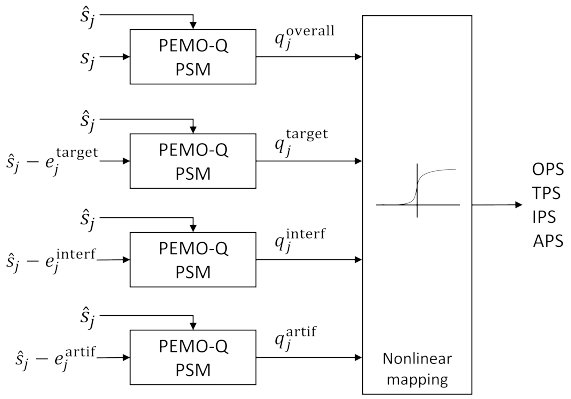


Figure 2.1: Block diagram for the computation of objective measures OPS, TPS, IPS, and APS (Reproduced from [14])

where $e_{ij}^{\text{target}}(t)$, $e_{ij}^{\text{interf}}(t)$, and $e_{ij}^{\text{artif}}(t)$ are distortion components due to target-distortion, interference, and artifacts respectively.

Prior to its decomposition into distortion components, both the estimate signal $\hat{s}_{ij}(t)$ and the target signal $s_{ij}(t)$, are processed through a gammatone filter bank whose center frequencies are linearly distributed on the Equivalent Rectangular Bandwith (ERB) scale. This allows for the distortion estimates to be weighted with respect to the loudness levels of various frequency bands, as defined by the gammatone filter bank.

Perceptual Similarity Measure (PSM) provided by the PEMO-Q auditory model [26] is used to compare the difference in the estimated signal and the distortion components as shown in figure 2.1 and Equations (2.15 - 2.18).

$$q_j^{\text{overall}} = \text{PSM}(\hat{s}_j, s_j) \tag{2.15}$$

$$q_j^{\text{target}} = \text{PSM}(\hat{s}_j, \hat{s}_j - e_j^{\text{target}}) \tag{2.16}$$

$$q_j^{\text{interf}} = \text{PSM}(\hat{s}_j, \hat{s}_j - e_j^{\text{interf}}) \tag{2.17}$$

$$q_j^{\text{artif}} = \text{PSM}(\hat{s}_j, \hat{s}_j - e_j^{\text{artif}}) \tag{2.18}$$

The PEASS objective evaluation measures the Overall Perceptual Score (OPS), Target-related Perceptual Score (TPS), Interference-related Perceptual Score (IPS), and Artifacts-related Perceptual Score (APS) are calculated by non-linearly mapping

the features q_j^{overall}, q_j^{target}, q_j^{interf}, and q_j^{artif} to the results of a listening test performed by the authors, using a feed forward neural network with sigmoid basis functions such that the mean square error (MSE) between the objective score and the mean subjective score (MOS) is minimized [14].

Although the authors report high correlation between the subjective ratings and the PEASS measures, no other independent evaluation of the performance of these measures in comparison to subjective listening tests could be found at this time.

CHAPTER III

METHODS OF SUBJECTIVE DATA COLLECTION

In order to determine the performance of objective measures (for singing voice separation) in comparison to human evaluation, a dataset containing the subjective ratings for audio processed with SVS techniques is necessary. The lack of availability of such a dataset necessitated an effort towards development of a new dataset. The subjective data collection was done in parts using three different listening experiments.

The initial listening-experiment (LE1) was performed in a controlled environment using identical equipment and environment for all participants in the study. The purpose for this was to serve as a tool for preliminary analysis and served two main purposes. Firstly, the subjective ratings obtained from LE1 were used as ground truth for evaluation of the SVS algorithms and thus were used to measure the accuracy of the current-state-of-the-art objective measures [21]. These results are discussed in Chapter 5. The second use for the subjective ratings from this listening experiment was to use them as a training set for development of new objective measure. Separate subjective ratings were necessary in order to test and validate the performance of the newly proposed objective measures. To obtain the testing and validation datasets two new listening experiments (LE2 and LE3) were conducted. During the design phase of the three listening experiments, LE1, LE2, and LE3 it was ensured that training and testing data was disparate in audio-content, SVS techniques, and participants in order to obtain an unbiased estimate of performance for the new objective measures.

The remainder of this chapter details the methods and protocols used for obtaining subjective ratings for SVS evaluation and discusses the philosophy behind the design of listening experiments. The procedures used for cleaning up the noisy data and a

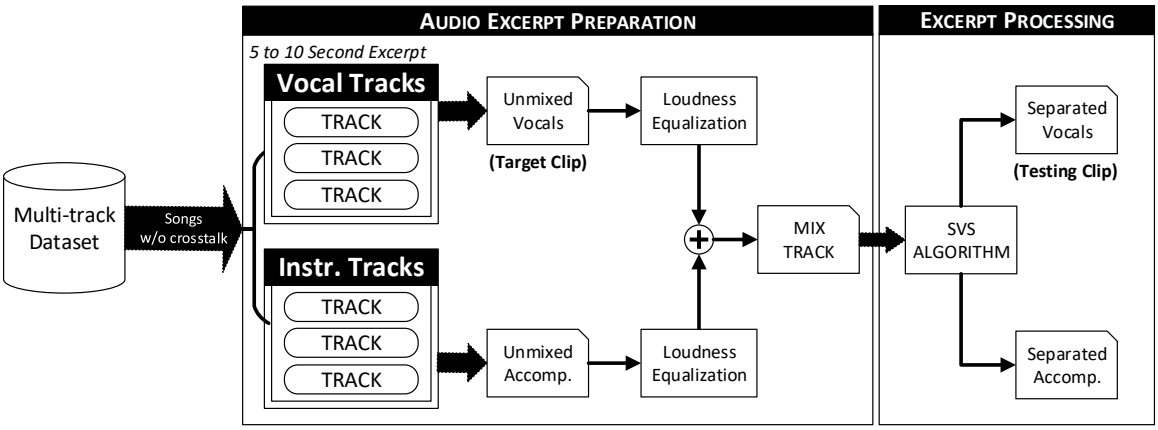


Figure 3.1: The process used to prepare an excerpt from a song for use in the listening experiments.

detailed analysis of the perceptual rating obtained from this effort are also presented in this chapter.

3.1 Design Considerations

The objective behind all three listening experiments was to compare and rate the vocals extracted from the mixed audio using various SVS techniques in a subjective assessment. A similar design process was followed for each of them as described below.

To prepare the audio data, an excerpt from a song was processed through different SVS algorithms. The participants were then asked to listen to the vocals extracted by these algorithms and compare them based on different judgment criteria in a series of tasks. This process was repeated with a selection of excerpts from multiple songs and each participant was asked to listen and rate a randomly sampled subset of the song choices in the experiment. This process is demonstrated visually in Figure 3.1.

The comparison of extracted vocals by the participants in the experiments was performed using a variation of the ITU-R BS.1534 MUSHRA standard [32, 53]. ITU-R BS.1534 MUSHRA listening test protocol is designed to compare multiple identically sourced audio clips with moderate to severe impairments. According to this protocol the subject is provided with an audio clip which is marked as the target or reference,

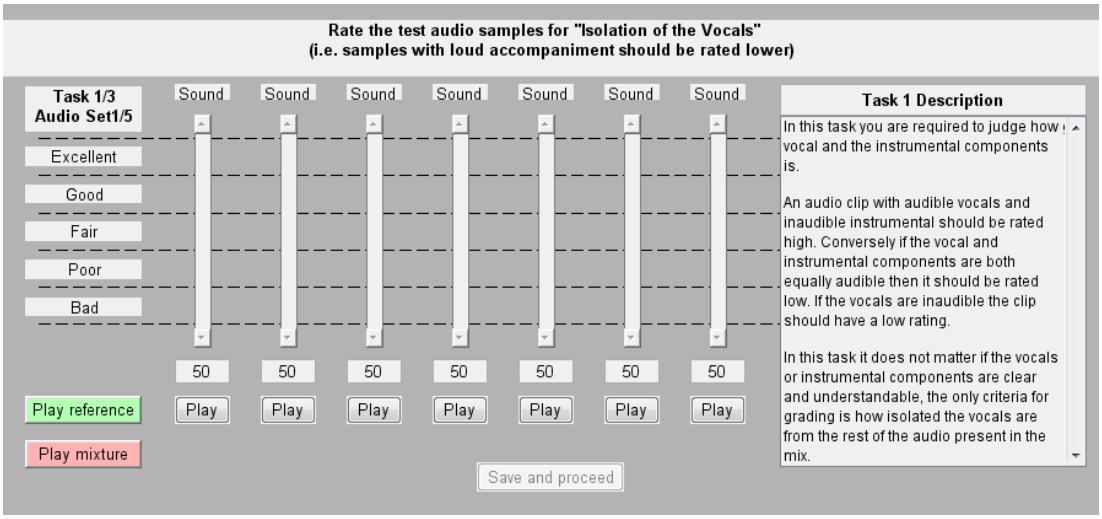


Figure 3.2: The MUSHRA interface used in the first listening experiment (LE1).

and a testing set consisting of multiple audio clips which are a result of processing the reference audio using different SVS algorithms. The subjects are expected to rate the clips in the testing set in terms of closeness to the target clip. Control points are added to the testing set in the form of artificially degraded anchor clips (expected to perform the worst), and a hidden reference clip (identical to the known reference; expected to perform the best). The participants are provided with a graphical user-interface (GUI) on a computer screen which allows them to listen to the reference clip along with the clips in the testing set, and rate them using slider controls. Each participant is required to rate the testing set clips on a scale of 0 to 100, with equidistant markings providing labels *Bad, Poor, Fair, Good* and *Excellent* going in increasing order. The closer a testing clip is to the target, the higher it should be rated. The GUI which was used for the MUSHRA test for LE1 is shown in Figure 3.2.

The design decisions for the three experiments and the reasons behind them are explained below. (A detailed report specific choices for audio content and SVS algorithms used to process them is available in Appendix A).

3.1.1 Choice of Audio Excerpts

For comparing the results of subjective evaluation against the objective measures it is necessary that the unmixed vocal and instrumental tracks are available. To this end publicly available corpora of multi-track music recordings were used to procure the excerpts for the listening experiments. Excerpts of five to ten seconds in duration, containing both vocal and instrumental components, were sampled from songs in the MedleyDB database [3] and "The Open Multitrack Testbed (OMT)" [9]. The song choices for the listening experiments were randomly chosen subsamples of all the available song choices in multi-track corpora. It was ensured that the track recordings for the chosen songs had no cross-talk between the vocal and instrumental tracks so that the ground truth samples of perfectly separated tracks were available. The test cases were generated by mixing the vocal and accompaniment tracks after normalizing them to equal loudness using Zwicker's loudness (ISO 532B) approximation [45, 69].

Nine excerpts were chosen only from the MedleyDB dataset in the case of LE1 and each participant was asked to rate five of them. For LE2 however, six excerpts each were chosen from each of the two corpora (MedleyDB and OMT). Each participant was asked to label only two excerpts as the number of evaluation tasks was increased from three to five (discussed below in Section 3.1.3). It was ensured that song choices did not overlap between LE1 and LE2. In the case of LE3, the choice of excerpts was a superset of the excerpts used in LE2. In addition to the twelve excerpts used in LE2, three excerpts each were added from both the corpora in LE3 bringing the total to eighteen excerpts. The participants in LE3 were paid volunteers (as described in Section 3.1.2), and therefore asked to rate three excerpts each.

3.1.2 Choice of Participants

The selection of participants is an important factor for any subjective study. While the original MUSHRA protocol calls for the participation of expert listeners, such

constraints were removed from the choice of subject pool due to the lack of availability of trained listeners. For LE1 the subjects were gathered from a normal hearing population of graduate and undergraduate students, with ages varying from nineteen to thirty-six. Out of thirty subjects who participated, eleven had experience in a music related field and six were professionally trained in music and/or had studio recording experience. The others were not trained in music.

The pool of participants for LE2 and LE3 however was more diverse due to the use of a web based system instead of a controlled listening environment. It was observed from the processed audio excerpts used in LE1 that the audio impairments present in the extracted vocals using existing SVS methods are severe enough that it may not have been necessary to use identical equipment or controlled testing environment. This observation was also corroborated by Cartwright et al. [6] in which they compared the results of two source separation based listening experiments with identical content. One of the listening experiments was conducted under controlled conditions and the other one was crowd-sourced using Amazon Mechanical Turk. Their findings showed that there was no statistically significant difference in the results obtained from the two listening experiments.

While the lack of controlled environment may result in an increase of variance in the ratings, the advantage of this approach is that sample size of population for the experiment is increased. For LE2 the participants consisted of volunteers recruited using word of mouth, social networks, and advertising on DSP and MIR related communities and varied in ages from nineteen to sixty-nine years old. Out of ninety-one participants who completed the listening experiment twenty-five identified that they had no prior experience with listening experiments while sixty-four reported to have participated in similar studies before. Unlike the other two experiments which were based on voluntary participation of interested populace, the subjects in LE3 were paid workers recruited using Amazon's Mechanical Turk service. In this case, out of

one hundred and forty-five participants, only ninety-two completed the listening-test with sixty-four self-reporting at least some prior experience. These participants varied in ages from seventeen to sixty-nine.

For all the three cases, there was no pre-screening of participants performed so as to keep the duration of the test and participant fatigue low; however, ratings from subjects which did not rank the hidden-target and the anchor clips appropriately were filtered out during post-processing analysis of the experiments (as explained in Section 3.2). Additionally, no diagnosis or test was performed prior to the listening test to determine if a participants' auditory perception was typical or not.

3.1.3 Judgment Tasks

The separated vocals produced using the different SVS algorithms were evaluated by the participants of the listening-test across five tasks based on different criteria. These tasks were —

T1 **Overall Quality:** The participants were asked to judge their overall perception of the quality of the separation, taking all impairments into account. The purpose here was to determine what factors influence the perception of separation of vocals for an average listener in a comprehensive sense, and if different listeners place importance on different factors which may be important for evaluating separation of vocals in music.

T2 **Vocal Isolation:** The participants were asked to judge how well-isolated the vocals were from the accompaniment. This task was designed to evaluate the ability of the algorithms to remove interference from other sources (instrumental accompaniment) when separating the vocals.

T3 **Target Preservation:** The participants were asked to judge how well the vocals were preserved in the testing samples as compared to the target. The

intent behind this task was to get a perceptual measure of subtractive distortion. In SVS techniques subtractive distortion of the target may happen when the vocal portion of the audio may partially or fully removed when attempting to remove the instrumental accompaniment.

T4 **Absence of Artifact Noise:** The participants were asked to judge how much noise was added in form of processing artifacts to the separated vocals in comparison to the target. In this case, the task was designed to judge the performance of the SVS algorithms at separating the vocals without additive distortion in the form of perceivable artifacts being introduced in the resulting audio.

T5 **Preservation of Intelligibility:** The participants were asked to judge how well the intelligibility of the vocals was preserved after separation as compared to the reference (original clip). Intelligibility is an important characteristic of speech, and this task was devised to determine if the process of separation reduced the perceived intelligibility of the separated vocals as compared to the source material.

For the preliminary experiment (LE1), only the Overall Quality, Vocal Isolation and Preservation of Intelligibility tasks (T1, T2, and T5) were used. The remaining two tasks were added for the later experiments based on the feedback provided by the signal processing community in response to the publication of the results of LE1 in [21].

3.1.4 Anchor Clips

As stated earlier in the chapter, artificially degraded test clips called anchors along with a hidden reference are used as a part of MUSHRA protocol. The purpose of the anchors in the experiments is to allow for screening subjects who may not have

understood the task or may have anomalous ratings for other reasons. The anchors are designed such that depending on the task, the subject will provide a very low score to the anchor clip if the task is being performed correctly. The anchor clips that were incorporated in these experiments are -

A1 **Interference Anchor:** The audio clip to be used as the anchor for the Vocal Isolation Task (T2) is designed such that the audio quality is degraded without any changes in the relative loudness between the target (vocals) and interfering components (instrumentals). To this end, the interference anchor is constructed by passing the original excerpt (mix of vocal and instrumental) through a 4 KHz low-pass filter, and amplifying the result to match the original loudness. This anchor is expected to rank the lowest among the test samples for the Vocal Isolation Task.

A2 **Subtractive Distortion Anchor:** This anchor is generated by passing the clean vocal audio to a 500 Hz low-pass filter. The result is then amplified such that the average Zwicker's Loudness (ISO 532B) for the result, is equal to the loudness of the original audio sample [45, 69]. The subtractive distortion anchor clips are the distorted vocals remaining after aggressively removing high frequency components from the target. As a result, this anchor is expected to be ranked the lowest in the Target Preservation Task (T3).

A3 **Additive Distortion Anchor:** In this case, the attempt is to simulate the kind of additive noise which is produced as processing artifacts by the current state-of-the-art singing voice separation techniques. To this end, the additive distortion anchor clips are generated by decomposing the unmixed vocals into a time-frequency representation using short-time Fourier transform (STFT) and randomly masking a portion of these time-frequency coefficients. The resulting signal is converted back to time domain representation using inverse STFT.

This signal is then mixed with the unmixed vocals to get the anchor clip. Prior to their use in the listening experiments these anchor clips are adjusted to have the same average loudness as the target audio. This anchor is expected to be ranked the lowest in the Absence of Artifact Noise Task (T4).

For the Preservation of Intelligibility Task (T5), any of the two distortion anchors (A2 and A3) may be expected to score the lowest. Any ratings for Task (T5) may be considered eligible if either of these two anchors take the lowest rank, provided that they meet all other criteria for inclusion. Similarly, any of the three anchors if rated the lowest in the Overall Quality Task (T1) should indicate a valid rating if all other criteria are satisfied. Table 3.1 lists the anchors associated with each of the judgment tasks of the listening experiment.

Although the design procedure described above was used for the listening experiments, the three experiments differed from each other in the specifics. The implementation details for all the three experiments are provided in Appendix A.

3.2 Post-Processing Methodology

The listening experiments described in the previous section involve a subjective study of human perception. This necessitates that a post-test screening be performed where the ratings for the subjects who may not have understood a particular task, or who may be outliers in the group, are removed. The post-processing steps involved in screening the subjective ratings are described below.

Table 3.1: Anchors associated with listening experiment tasks

Task		**Anchor(s)**
T1	Overall Quality	A1, A2, or A3
T2	Vocal Isolation	A1
T3	Target Preservation	A2
T4	Absence of Artifact Noise	A3
T5	Preservation of Intelligibility	A2 or A3

3.2.1 Removal of Invalid Ratings

The ranks of the hidden reference and the anchor clip associated with each of the tasks is used to determine whether or not a subject has understood the judgment criteria for the task. To remove the ratings for subjects who may not have understood the task, the ratings that do not have the hidden reference scored the highest are removed. Additionally, for each of the tasks the ranking of their associated anchor clip(s) is determined. The ratings are treated as incongruous and removed from the study if (one of) the anchor clip(s) is not scored as performing the worst.

3.2.2 Removal of Outliers

After the incongruous subject ratings have been removed from the acquired dataset, the next step is to remove the ratings which may be outliers to the average subject agreement results. This process is performed by finding the Spearman's correlation coefficient (ρ) of the individual subjective ratings versus the mean of the ratings all the other subjects. Spearman's correlation coefficient represents the similarity in the rank-order of the two series being compared. The value of the coefficient lies between negative and positive one, with positive one indicating full agreement in terms of rank-order while negative one implies the that the two series have completely opposite rank-orders [57]. From the ρ values obtained by the process above, it may be inferred that the subjects which have low ρ values have provided ratings contrary to the general consensus and therefore must be removed from the pool or ratings as outliers. A truncated t-distribution is fitted over the set of ρ values obtained, and the subject ratings which have ρ values less than the five percent outlier limit on the lower tail of the distribution are removed [58]. This process is repeated for each pair of tasks and the test-sets of audio-excerpts.

3.2.3 Improving Inter-Rater Reliability

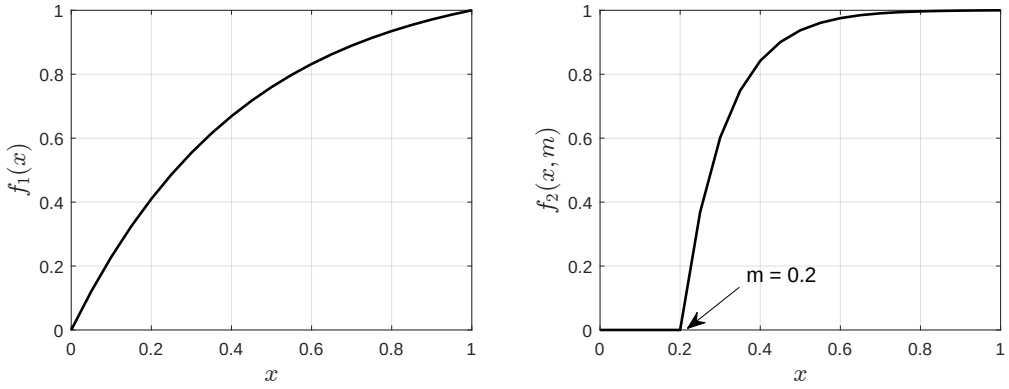
The next step in the process of post-evaluation analysis is determining the reliability of the ratings for each of the tasks. To ensure that the subject ratings are concordant with each other and the results are reliable, inter-rater reliability is determined for each task using Krippendorff's alpha coefficient [22]. Krippendorff's alpha (KA) is a statistical tool used as measure of the agreement between multiple ratings on identical samples, which can be used to compare two or more ratings on nominal, ordinal, or numeric scales. This makes KA a useful resource to estimate the reliability of the ratings obtained in the listening experiments discussed here as there are multiple numeric ratings for each task and excerpt which need to be compared. KA values lie between zero and one, with greater agreement being characterized by values closer to one.

In order to improve the inter-rater reliability of the results the remainder of the ratings, after the removal of outliers, are sub-sampled using a genetic optimization algorithm as described below. The objective of this genetic optimization is to find the subset of ratings which maximizes the KA value for a given judgment task without removing too many subjective ratings from the set. The process begins with defining two functions f_1 and f_2 as shown in Equations 3.1 and 3.2.

$$f_1(x) := \left(\frac{1-10^{-Ax}}{1-10^{-A}}\right) \quad \forall\, x \in [0,1] \tag{3.1}$$

$$f_2(x,m) := U(x-m)\left(\frac{1-10^{-B(x-m)}}{1-10^{-B(1-m)}}\right) \quad \forall\, x \in [m,1] \tag{3.2}$$

Here $U(x)$ is the unit step function, and m is a number between zero and one. A and B are arbitrary control parameters which can be used to control the gradient of functions and were chosen to be 1.0 and 4.0 respectively in the current implementation. Sample curves for both Equations 3.1 and 3.2 are shown in Figure 3.3. For the purpose of plotting the curves the m value was chosen as 0.2.



(a) Sample plot of f_1 for $A = 1.0$ (b) Sample plot of f_2 for $B = 4.0$ and $m = 0.2$

Figure 3.3: Component functions used to derive the objective function for genetic algorithm based optimization.

Further define N, N_{min}, N_{total} respectively as the number of ratings in the current set for each iteration, the minimum permissible number of ratings, and the total number of subjective ratings present in the set after removing the outliers. The objective function that has to be maximized can then be written in terms of the KA value of the current set $\boldsymbol{\alpha}$ and N as shown in Eqn. (3.3).

$$\mathfrak{F}(\boldsymbol{\alpha}, N) := f_1(\boldsymbol{\alpha})^{W_\alpha} \times f_2\left(\frac{N}{N_{total}}, \frac{N_{min}}{N_{total}}\right)^{1-W_\alpha} \tag{3.3}$$

The objective function defined in Eqn. (3.3) is the weighted geometric means of the two terms (with weight of the KA sensitive term as W_α) is used to determine the set of the ratings for any particular audio excerpt and judgment task for which the KA value is maximized. The contour plot for the objective function as well as the direction of steepest ascent for various points is shown in Figure 3.4.

The optimization is performed iteratively by mutating a subset of the subjective ratings as described in Algorithm 3.1. After each iteration, the value of the objective function is calculated for the mutated set and if the value is higher than that of the previous iteration the mutation is retained otherwise discarded till there is no

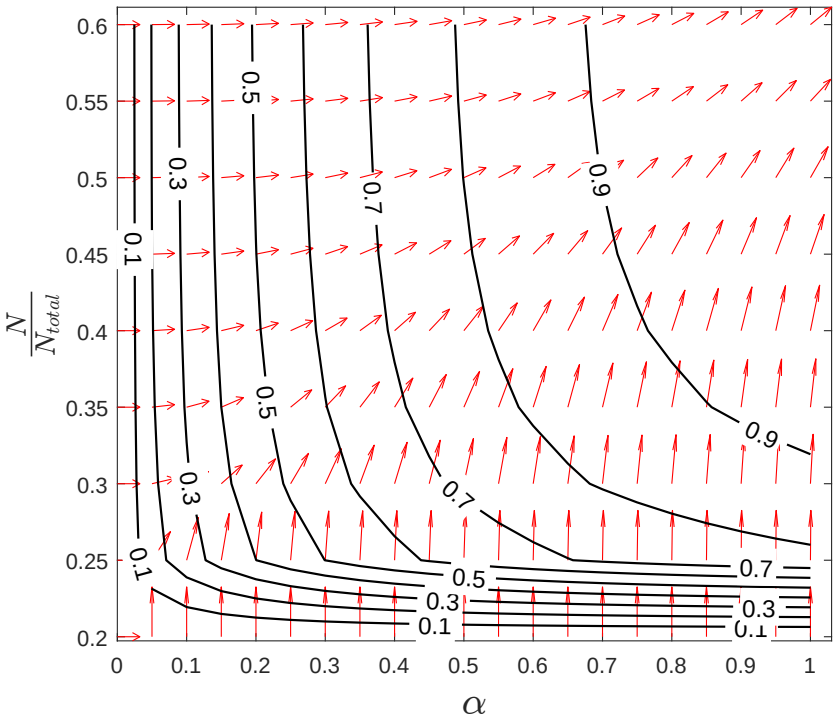


Figure 3.4: Contour plot of objective function for maximizing inter-rater reliability. W_α is set to 0.75 and N_{min}/N_{total} is 0.2. The arrows show the magnitude and direction of gradient of steepest ascent.

improvement for subsequent iterations. Since this algorithm converges to local-maxima, the process is repeated multiple times and the turn which provides the highest final value for the objective function is chosen.

This process of maximizing the inter-rater reliability was performed only for LE2 and LE3 as the KA values for LE1 were high enough that it was not needed.

Table 3.2 shows the results of the post-processing for all the three experiments in terms of the number of subjective ratings removed in each step and the final number of ratings used for final analysis in each case. The detailed analysis of each of the listening tests is provided in Appendix A.

Algorithm 3.1 Select subset of subjective ratings for maximizing KA

Initialize a random seed
Define: $I_{max} \leftarrow$ Maximum Iterations; $I_{min} \leftarrow$ Minimum Iterations; $\varepsilon \leftarrow$ Convergence Tolerance
$\mathcal{S} = \{1, 2, \cdots, N_{total}\}$
$\mathcal{S}_s = \text{RandSample}(\mathcal{S}, \lfloor 0.75 \times N_{total} \rfloor)$
$\mathcal{S}_c = \mathcal{S} - \mathcal{S}_s$
$\alpha_0 = \text{KripAlpha}(\mathcal{S}_s)$
$N_0 = n(\mathcal{S}_s)$
$V_0 = \mathfrak{F}(\alpha_0, N_0)$ % Objective function
for i from 1 to I_{max} **do**
  if i is even OR $N_{i-1} = N_{min}$ **then**
    $\mathcal{S}_{temp} = \mathcal{S}_s \bigcup \text{RandSample}(\mathcal{S}_c, 1)$ % Add random sample
    $\alpha_i = \text{KripAlpha}(\mathcal{S}_{temp})$
    $N_i = n(\mathcal{S}_{temp})$
    $V_i = \mathfrak{F}(\alpha_i, N_i)$
    if $V_i \geq V_{i-1}$ **then**
      $\mathcal{S}_s = \mathcal{S}_{temp}$
      $\mathcal{S}_c = \mathcal{S} - \mathcal{S}_s$
    else
      $V_i = V_{i-1}$
      $N_i = N_{i-1}$
    end if
  else if $N_{i-1} < N_{total}$ **then**
    $\mathcal{S}_{temp} = \text{RandSample}(\mathcal{S}_s, N_{i-1} - 1)$ % Remove random sample
    $\alpha_i = \text{KripAlpha}(\mathcal{S}_{temp})$
    $N_i = n(\mathcal{S}_{temp})$
    $V_i = \mathfrak{F}(\alpha_i, N_i)$
    if $V_i \geq V_{i-1}$ **then**
      $\mathcal{S}_s = \mathcal{S}_{temp}$
      $\mathcal{S}_c = \mathcal{S} - \mathcal{S}_s$
    else
      $V_i = V_{i-1}$
      $N_i = N_{i-1}$
    end if
  end if
  if $i > I_{min}$ **then**
    $M = \text{mean}(V_{i-4}, V_{i-3}, \cdots, V_i)$
    if $|V_j - M| \leq \varepsilon$ **for all** j **in** $i-4, i-3, \cdots, 1$ **then**
      break for % Convergence
    end if
  end if
end for
$\mathcal{S}_s$ is the subset which maximizes inter-rater agreement

Table 3.2: Number of subjective ratings remaining and removed after each post-processing step. The number of available ratings are shown in black, and the number of ratings removed in each step is in gray.

	LE1			**LE2**					**LE3**				
	T1	**T2**	**T5**	**T1**	**T2**	**T3**	**T4**	**T5**	**T1**	**T2**	**T3**	**T4**	**T5**
Initial Ratings #	142	142	142	167	155	155	155	139	346	326	346	322	315
Congruent Ratings #	2 144	-8 134	-30 112	-50 117	-85 70	-65 90	-68 87	-50 89	-127 219	-218 108	-172 174	-192 130	-155 160
Non-Outliers #	-22 **122**	-76 **58**	-25 **87**	-4 113	-8 62	-4 86	-10 77	-3 86	-8 211	-8 100	-17 157	-3 127	-11 149
Optimum KA #	Not Applicable			-39 **74**	-3 **59**	-11 **75**	-9 **68**	-12 **74**	-83 **128**	-12 **88**	-28 **129**	-14 **113**	-29 **120**

CHAPTER IV

ANALYSIS OF SUBJECTIVE RESULTS

In Chapter 3, the methodology used for three listening experiments was discussed. This chapter discusses the results obtained from the three experiments LE1, LE2, and LE3 as well as compares the data gathered from them. In Section 4.1 intra-task analysis is discussed and the reliability of the experimental results is determined. Section 4.2 compares the similarity or dissimilarity between the tasks of the same experiment and explores how ratings from perceptual judgment tasks correlate with each other. The chapter concludes with a discussion about the repeatability of the results obtained from the listening experiments in Section 4.3 and illustrates whether the tasks are performed consistently and reliably.

4.1 Reliability of Experimental Results

The reliability of the experimental results is dependent upon the inter-rater agreement of the listening experiment tasks. The data obtained from the listening experiments can be considered dependable if the subjects are consistent in assigning ratings and the subjective opinion of a majority of the raters agree to form a common judgment. This consistency of judgment is approximated using Krippendorff's Alpha (KA) [22] measure of inter-rater reliability as described in Chapter 3, Section 3.2.3.

Figure 4.1 shows the KA values for the judgment tasks from the listening experiments LE1, LE2 and LE3. Inter-rater agreement using Krippendorff's Alpha coefficient is calculated for each audio excerpt in the listening set. The figure shows the median, minimum, and maximum KA values across all the excerpts for different tasks.

In the case of LE1, the median of KA values for agreement among the subjects for

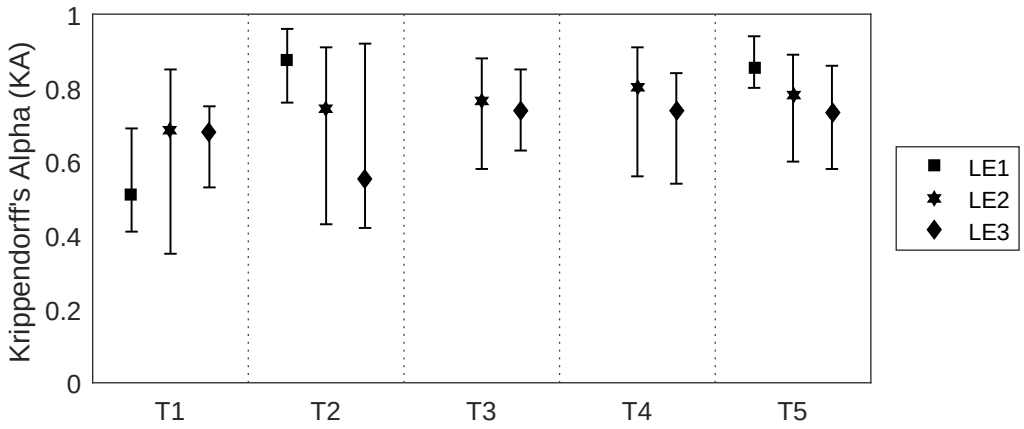


Figure 4.1: Inter-rater agreement for listening experiments measured using Krippendorff's Alpha (KA) for all the tasks in the three experiments. The marker positions indicate the median value and the error-bars show the maximum and minimum values obtained from calculating KA for different excerpts.

each of the excerpts, for the Vocal Isolation (T2) and Preservation of Intelligibility (T5) tasks was measured as 0.88 and 0.86 respectively. The median of KA values for the Overall Quality Task (T1) is however only 0.51. These results show that while the raters largely agreed with each other when it came to the tasks involving judging based on a specific category (T2 and T5), they could not form a consensus when asked to provide overall judgment based on multiple factors. The other two tasks were not a part of LE1.

Similar trends in inter-rater agreement were seen for LE2 and LE3. The median KA values for the Overall Quality Task (T1) in LE1 and LE2 were measured as 0.69 and 0.68 for LE1 and LE2 respectively. The median KA values for other tasks for both LE1 and LE2 were in the range of 0.74 to 0.80 (except T2 for LE2)which indicates that the subjects were more in agreement with each other.

It can be argued that a lack of agreement among the raters for any subjective study will have a detrimental affect on its use as a basis for comparison for other measures. Flexer and Grill have studied this effect [16] and demonstrate that limitations on inter-rater agreement forces an upper bound on the evaluation of MIR systems such as

"Symbolic Melodic Similarity" or "Audio Genre Classification". This also holds true for evaluation of SVS objective measures and therefore conclusions about performance of the objective measures should only be based upon the subjective results which demonstrate a high rater reliability.

Krippendorff's Alpha is a statistical measure, that basically is a measurement of noise in the data and can not be interpreted in an absolute sense. Therefore, it is difficult to determine a threshold for KA, above which it can be demonstrated that the reliability is statistically significant. Krippendorff however, remarks that a reasonable value of α for high reliability between two subjective variables is α greater than 0.8; for values of α between 0.67 and 0.8 the conclusions drawn can only be tentative [39].

The discussion above demonstrates that the subjective ratings obtained in LE1 for the vocal isolation task and preservation of intelligibility task (T2 and T5) are of good enough quality to be used as the basis of comparison for objective measures. However, the reliability of the ratings from the overall quality task (T1) is low and thus would not form an adequate source for comparison. Due to these results, two new objective measures have been proposed in this thesis that are modeled to emulate the ratings obtained for the vocal isolation task and preservation of intelligibility task (T2 and T5) of LE1 as discussed in Chapters 6 and 7. Although the latter two experiments have expanded the set of tasks to include the preservation of vocals and the absence of artifacts tasks (T3 and T4) and the reliability of the ratings are moderate, their main use in the present discussion pertains only as tools for evaluating the current state-of-the-art objective measures. Designing new objective measures for these criteria will be left as an endeavor for the future.

4.2 Comparison of Evaluation-Tasks

The subjective assessment for the SVS separated vocals was evaluated for five different tasks, each having a particular judgment criteria, as discussed in Chapter 3. In this

section, the ratings obtained from these tasks are compared against each other in a pairwise manner in order to determine if the subjective evaluation for each task is independent of the other tasks.

The evaluation procedure in this case starts with the normalization of the MUSHRA [32] ratings from all the experiments by subtracting the means of the ratings and dividing by their standard deviations as shown in Eqn. (4.1).

$$\acute{r}_{sea} = \frac{r_{sea} - \mu_{se}}{\sigma_{se}} \tag{4.1}$$

Here $\mathbf{r_{sea}}$ is the rating provided by subject **s** for audio excerpt **e** which has been processed by SVS algorithm **a**; $\mu_{\mathbf{se}}$ and $\sigma_{\mathbf{se}}$ are the mean and standard deviation of the ratings across all SVS algorithms for **s** and **e**.

The normalized ratings ($\acute{\mathbf{r}}_{\mathbf{sea}}$) from each subject were averaged to get the final rating for each audio except as shown in Eqn. (4.2). This process was repeated for all the judgment tasks which resulted in all the audio clips being assigned a single representative subjective-score ($\hat{\mathbf{r}}_{\mathbf{ea}}$) for each of the tasks.

$$\hat{r}_{ea} = \frac{1}{S} \sum_{s} \acute{r}_{sea} \tag{4.2}$$

Spearman's correlation coefficient [57] was calculated for each of the excerpts between the normalized ratings for every pair of tasks. Table 4.1 shows the results of this analysis in the form of average Spearman's correlation between each pair of tasks (ratings from all three listening experiments are averaged).

The results in Table 4.1 can be interpreted in a number of ways that suggest various possibilities for the underlying structure of the data. The ratings of the intelligibility preservation task (T5)demonstrate a strong positive correlation versus the vocal target preservation (T3) and artifact noise (T4) tasks. Therefore, one possible inference from this analysis that intelligibility in the separated vocals is in a large part adversely affected by the presence of artifacts and subtractive distortion of

Table 4.1: Pairwise Spearman's correlation between ratings of judgment tasks averaged across listening experiments

T2	**T3**	**T4**	**T5**	
-0.08	0.33	0.45	0.47	**T1**
	-0.50	-0.44	-0.50	**T2**
		0.78	0.85	**T3**
			0.74	**T4**

the audio.

Similarly, tasks T3 and T4 are themselves strongly correlated in their ratings and two possible theories may be able to explain this phenomenon. Firstly, the high correlation may be due to the mechanics of SVS algorithms where introduction of additive and subtractive distortions in the resulting audio are not mutually independent processes. This explanation is also supported by the negative correlations between vocal isolation task (T2) and tasks T3 and T4, which means more aggressive an algorithm is at isolating the vocal component from the audio, more distortions it introduces. Alternatively, the cause for high correlation between T3 and T4 may be that the participants in the listening tests are not expert listeners and consequently are not able to adequately discern between additive and subtractive distortion.

The usefulness of the analysis presented in this section is that it provided a starting point for the development of the objective measure for estimating the preservation of intelligibility in separated vocals as depicted in Chapter 7.

4.3 Inter-experiment Agreement

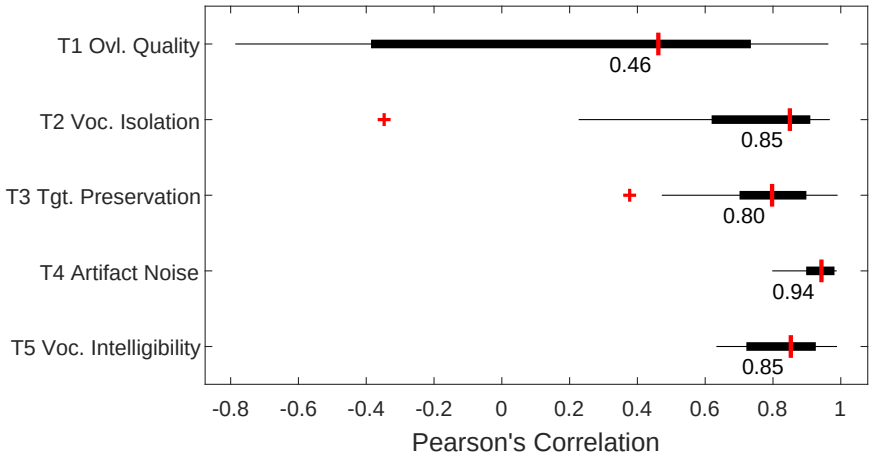
In this section, the results obtained from different listening experiments are compared to determine whether there is agreement among them. The question of interest in this case is if the subjective assessment of two different groups of raters is in agreement when they are rating the same audio samples. Unfortunately, because of the design choices, this analysis can only performed between LE2 and LE3 as some audio content

is common between them which is not the case with LE1. As described in Chapter 3, this design choice was made so as to provide a substantial differences in the content of the training and testing datasets, which is important to test the merits of any objective measures developed using the results from LE1.

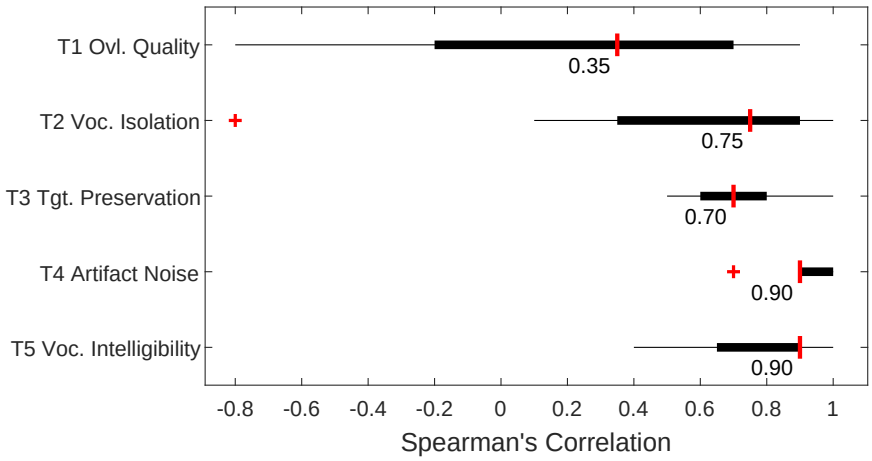
In order to compare the ratings from LE2 and LE3, the representative scores for each task (as described in Section 4.2) were once again used. For the representative score from each of the tasks in one of the listening experiments its Pearson's correlation coefficient (PCC) [47] and Spearman's correlation coefficient (SCC) [57] with the corresponding task from the other experiment was calculated. These quantities respectively estimate whether or not two variables being compared have a linear relationship and similar rank-order structure. The process was repeated for all the audio excerpts which are common to both LE2 and LE3 and the results are shown in Figure 4.2.

The box and whisker plot in Figure 4.2 shows that for the same audio content, subjective opinion of two different sets of raters agrees to a high degree for linear and rank-order of performance of different SVS algorithms in the case of tasks T3, T4 and T5. For these three tasks the median PCC values (0.80, 0.94, and 0.85) as well as median SCC values (0.70, 0.90, 0.90) are all very high with the distribution of values tightly concentrated (as evidenced by the small whisker spans). This result indicates that the tasks were well designed and defined to the effect that they provide repeatable results, which is an essential characteristic for any experimental study.

The correlations in case of task T2, while having high median values, suffer from having a wider distribution. The underlying factor for this result is the poor inter-rater agreement among the raters in LE3 for task T2 as discussed in Section 4.1. The heavy skew of the medians towards higher values in this case suggests that the large spread in the correlation values is an artifact of disagreement in the ratings for a small subset of the audio excerpts common to both LE2 and LE3. The analysis is



(a) Box Plot of Pearson's Correlation (PCC) Values



(b) Box Plot of Spearman's Correlation (SCC) Values

Figure 4.2: Comparison between LE2 and LE3 based on the audio excerpts common in both the datasets. Median values are marked as red lines with actual values written next to them.

unclear on weather the low correlation values in task T2 for some of the excerpts is a chance occurrence, or they are an indication of a design flaw in this task that affects the repeatability of results adversely.

For task T1, as shown in Section 4.1, there is very little agreement between the raters for the same listening experiment for both LE2 and LE3. This suggests that the normalized average scores are not really representative of the subjective opinions of the participants and a lack of consensus among the participants makes it ineligible

for use in further analysis.

CHAPTER V

ASSESSMENT OF OBJECTIVE MEASURES

In chapter 2 objective measures such as BSS_Eval measures and PEASS measures which are the current state-of-the-art for automatic evaluation of source separation techniques were introduced. BSS_Eval Measures comprising of SIR, SAR and (N)SDR were developed to estimate the performance of general source separation algorithms and they have been widely used of evaluation of singing voice separation task in competitions like Music Information Retrieval and Exchange (MIREX) [10, 11, 8]. In contrast, PEASS Measures (OPA, IPS, TPS, and APS) have been optimized specifically for evaluation of auditory source separation tasks like the cocktail party problem or even singing voice separation, but haven't been widely adopted by the community.

In this chapter, BSS_Eval and PEASS measures are evaluated in the context of singing voice separation by comparing the objective scores to the subjective-assessment ratings provided by the listening experiments as discussed in Chapter 3.

The performance of the objective measures is compared to the subjective ratings for two characteristics, linear dependence and monotonicity. Linear dependence is measured in this case by Pearson's Product Moment Correlation Coefficient (PCC) [47]. A high linear dependence between objective scores and subjective ratings implies that the objective measure being evaluated shows a low amount of variance with respect to the subjective ratings. Monotonicity is evaluated using Spearman's Rank-Order Correlation Coefficient (SCC), which is a measure of ordinal relationship between two variables [57]. A high SCC value implies that an increase or decrease in the subjective rating is accompanied by a corresponding increase or decrease in

the objective measure result. Both PCC and SCC values lie in the range of negative one to positive one. A positive value for PCC or SCC between two random variables $\overrightarrow{X}$ and $\overrightarrow{Y}$ implies that a positive or negative change in $\overrightarrow{X}$ shall be accompanied by a change in $\overrightarrow{Y}$ in the same direction. However, a negative correlation coefficient is observed if the changes in $\overrightarrow{X}$ and $\overrightarrow{Y}$ are in opposite directions.

The steps taken to compare the performance of the objective measures against subjective ratings are listed below.

Audio Processing: An audio excerpt comprising of vocal and instrumental components is chosen and processed using multiple different SVS techniques to get several instances of separated-vocals.

Subjective Evaluation: The instances of separated-vocals are compared to each other using subjective listening tests as described in Chapter 3.

Objective Evaluation: Scores from various objective measures are calculated for each instance of separated-vocals.

Comparison: The corresponding subjective ratings and objective measure scores are compared against each other using PCC and SCC. The PCC and SCC values are calculated for all subjects' ratings.

The process described above is repeated for all the audio excerpts in the listening experiment to obtain a list of PCC and SCC values for each of the tasks and objective measures. In order for an objective measure to be determined as suitable for a given perceptual evaluation task, the average of correlation values obtained as described above should be high (either positive or negative). Additionally, the PCC or SCC values should be consistently positive or negative (but not both) for the objective measure to be considered reliable, and the correlation between the two quantities to be statistically significant. To determine the reliability of the objective

measure, the list of PCC or SCC values are fitted to a scaled and shifted student's t-distribution (truncated to a range of -1 to +1). The boundary limits for the central ninety-five percent area under the probability density function (PDF) are determined from the probability distribution estimate of the PCC/SCC values and used as the confidence interval for comparison. The objective measure can be said to demonstrate a statistically-significant correlation with the subjective assessment if the confidence interval lies entirely in the positive or negative sides of the number line.

5.1 Performance of BSS_Eval Measures

Table 5.1 shows the results of comparing the BSS_Eval Measures with the perceptual ratings from the relevant tasks in each of the listening experiments. Signal to Interference Ratio (SIR) is the log-energy ratio of the signal of interest (vocals) to the other interfering signals (accompaniment); it is therefore expected that the SIR values should show a positive correlation with the Vocal Isolation Task (T2). Similarly Signal to Artifacts Ratio (SAR) and Normalized Signal to Distortion Ratio (NSDR) measure the log-energy ratios between the target signal and artifacts noise and total distortion respectively. SAR, therefore has been compared to the ratings obtained in the preservation of absence of artifact noise task (T4), and NSDR against the overall perceptual quality rating task (T1). Additionally upon comparison, the results from various tasks in the listening experiments suggest that the perceived preservation of vocal intelligibility is largely dependent upon the amount of additive distortion present in the separated vocals. Therefore, in table 5.1 an additional comparison between SAR and T5 is also presented. Unfortunately, BSS_Eval measures do not have a specific quantity which can be used to measure the amount of subtractive distortion present in the target signal; therefore, none of these measures are compared against the vocal preservation task (T3). In the table, the results are presented in the form of the average PCC or SCC value along with the ninety-five percent confidence

Table 5.1: Average PCC and SCC values for subjective ratings v. the corresponding BSS_Eval measures with ninety-five percent confidence intervals. The instances which show a statistically significant correlation have been typeset in bold.

		T1 v. NSDR	**T2 v. SIR**	**T4 v. SAR**	**T5 v. SAR**
Pearson's Corr. Coefficient (PCC)	**LE1**	0.119 [-.90,+.96]	0.567 [-.65,+1.00]	N/A	**0.739 [+.07,+1.00]**
	LE2	0.351 [-.66,+.95]	0.614 [-.08,+.98]	0.516 [-.19,+.99]	0.466 [-.40,+.99]
	LE3	0.194 [-.67,+.96]	0.444 [-.62,+1.00]	0.480 [-.35,+.98]	0.439 [-.52,+.97]
Spearman's Corr. Coefficient (SCC)	**LE1**	0.348 [-.55,+.74]	0.701 [-.05,+.88]	N/A	**0.816 [+.32,+.93]**
	LE2	0.351 [-.66,+.95]	0.614 [-.08,+.98]	0.516 [-.19,+.99]	0.466 [-.40,+.99]
	LE3	0.212 [-.67,+.98]	0.383 [-.63,+.99]	0.506 [-.30,+.98]	0.403 [-.59,+.99]

interval in the brackets below them.[1]

The results obtained by comparing the subjective evaluational of SVS in the three listening experiments to the objective valuate using BSS_Eval measures shows that the objective evaluation is in general poorly correlated to the subjective ratings. The only statistically significant correlations between the subjective and objective results are obtained when comparing the intelligibility task ratings (T5) from LE1 with SAR, with average values of 0.74 and 0.82 for PCC and SCC respectively (highlighted in bold in table 5.1). However, this is not repeated in the other two listening experiments suggesting that this may be an isolated phenomenon and not a repeatable result.

While the results presented here do not conclusively prove that BSS_Eval measures may not have a good predictive value for these perceptual tasks, the evidence does

[1] An exhaustive comparison of each of the BSS_Eval measures to each of the tasks is provided in table B.1 on page 74.

suggest a high likelihood for this to be true. These results are in agreement with Emiya et al. [14], although in their case the evaluation was performed for source separation in general audio mixtures, only a subset of which constituted singing voice separation.

5.2 Performance of PEASS Measures

A similar analysis to the one described in the previous section has been performed using the PEASS Measures OPS, TPS, IPS, and APS and presented in table 5.2. The most relevant comparisons of interest in this case are OPS against the overall quality (T1), TPS versus the target preservation task (T3), IPS compared to vocal isolation ratings (T2), and APS against absence of artifact noise (T4). Again, the subjective ratings for preservation of intelligibility (T5) are compared to the objective ratings for the objective measure for additive distortion i.e. APS.[2]

The results from comparing the subjective ratings and the objective scores from PEASS measures demonstrate that APS and T4 ratings are significantly correlated, which indicates that APS provides consistent agreement with subjective observations for perception of added processing artifacts in the extracted vocals. In this case too, the APS measure shows statistical significance in correlation with the intelligibility ratings (T5) for LE1, but not with LE2 or LE3. However, the correlation between APS and T5 is stronger and has narrower confidence bounds than the correlation between SAR and T5.

The correlation between the vocal isolation ratings (T2) and IPS, while not being a statistically significant correlation, is however quite high for LE1 and LE2 (the low average correlation for IPS v. T2 in the third listening experiment might be explained by the poor inter-rater agreement for T2 in LE3). Both OPS and TPS fail to show any substantial correlation with the ratings from their respectively relevant subjective

[2]An exhaustive comparison of each of the PEASS measures to each of the tasks is provided in table B.2 on page 75.

Table 5.2: Average PCC and SCC values for subjective ratings v. corresponding PEASS measures with ninety-five percent confidence intervals. The instances which show a statistically significant correlation have been typeset in bold.

		T1 v. OPS	**T2 v. IPS**	**T3 v. TPS**	**T4 v. APS**	**T5 v. APS**
Pearson's Corr. Coefficient (PCC)	**LE1**	0.047 [-.83,+.88]	0.638 [-.38,+1.00]	N/A	N/A	**0.862 [+.35,+1.00]**
	LE2	0.247 [-0.79,+0.92]	0.633 [-0.20,+1.00]	0.207 [-0.80,+0.90]	**0.810 [+0.35,+1.00]**	0.593 [-0.10,+1.00]
	LE3	-0.009 [-0.87,+0.90]	0.386 [-0.80,+0.98]	0.098 [-0.72,+0.85]	**0.782 [+0.15,+1.00]**	0.612 [-0.13,+1.00]
Spearman's Corr. Coefficient (SCC)	**LE1**	0.012 [-.83,+.93]	0.638 [-.28,+1.00]	N/A	N/A	**0.797 [+.28,+1.00]**
	LE2	0.276 [-0.76,+0.95]	0.638 [-0.22,+1.00]	0.220 [-0.83,+0.92]	**0.783 [+0.23,+1.00]**	0.601 [-0.15,+1.00]
	LE3	0.069 [-0.81,+0.89]	0.379 [-0.71,+1.00]	0.137 [-0.67,+0.90]	0.711 [-0.02,+1.00]	0.560 [-0.24,+1.00]

assessment tasks.

Although comparisons between tables 5.1 and 5.2 indicate PEASS measures appear to in general perform better than BSS_Eval measures, PEASS measures have not found wide spread acceptance in the MIR and audio DSP communities. It is for this reason that both these sets of measures have been treated as the current state-of-the-art for objective evaluation of source separation in this thesis.

5.3 Need for New Measures

The great importance of robust techniques for Singing Voice Separation (SVS) in the field of Music Information Retrieval (MIR) can not be denied. Various algorithms for music information retrieval tasks such as melody extraction [12, 24, 67], automatic lyrics recognition [66], singer identification [17, 44], query by singing/humming [28], etc., use SVS or partial source separation techniques as one of the preprocessing steps

in their design. Unfortunately, as discussed in the previous sections, the objective measures (BSS_Eval and PEASS) used for assessing the quality of vocal separation provided by different algorithms have proved to be inadequate. It was demonstrated that the results of evaluating SVS techniques using these measures generally do not agree with the results from subjective evaluation. Additionally both sets of measures need the presence of unmixed vocal and instrumental tracks for evaluation. Outside of academic datasets, the unmixed tracks are rarely available, limiting the utility of these measures to evaluate SVS performance in real-world situations.

Two new objective measures are proposed in this thesis which overcome the limitations imposed by the current state-of-the-art. Vocal Isolation Score (VIS) is proposed as an objective measure to better evaluate the performance of various SVS implementations in context of isolation of the target in the separated vocals. The formulation and design of VIS has been described in Chapter 6. Another objective measure, Vocal Intelligibility-Preservation score (VIPS) has been designed to assess the separated vocals for preservation of perceptual vocal intelligibility and is described in Chapter 7. Unlike BSS_Eval or PEASS measures, both VIS and VIPS have been designed as "no-reference" measures, which means that they do not need the unmixed vocal and instrumental tracks for the evaluation.

CHAPTER VI

DEVELOPMENT OF ISOLATION MEASURE

In this chapter a new objective measure, Vocal Isolation Score (VIS), is proposed to evaluate the performance of SVS algorithms in the context of isolation of the vocals. The current processes for extraction of the vocals are not perfect and often result in the presence of interference in the form of residual accompaniment. VIS is designed to correlate with the loudness of this residual accompaniment signal such that extracted vocals with low residuals score better than ones which have more residuals and hence have worse vocal isolation. VIS is derived using regression from a set of features which are sensitive to the relative loudness levels of instrumentals and the vocals as discussed below in Section 6.1. The performance of VIS, for perceptual assessment of vocal isolation, as compared to the current state-of-the-art objective measures is evaluated and discussed in Section 6.2.

6.1 Design Process

Lehner et al. [40] demonstrated that for detecting regions of vocal activity in music, performance of appropriately parameterized Mel-Frequency Cepstral Coefficients (MFCC) is comparable to that of more complex features. Their results were independently verified as a part of this work using excerpts from MedleyDB Multitrack Dataset [3] and iKala Dataset [7] for both within dataset and cross-dataset training and testing. The feature extraction process is detailed in Section 6.1.1.

The results obtained above formed the motivation for investigating the perceptual significance of MFCC based features for assessment of vocal isolation from musical mixes. This investigation is described in Section 6.1.2, where the sensitivity of these features to the presence of instrumental music is explored. The last part of the process

is training a machine learning based regression model which combines these features to get the VIS. This process is described in Section 6.1.3

6.1.1 Feature Extraction

The feature extraction process begins with the normalization of extracted vocals with respect to loudness. The loudness of the extracted vocals is estimated using Zwicker Loudness Estimate (ISO 532B) [45, 69] and the signal is scaled such that the loudness of the normalized audio is sixteen sones. The normalized audio signal is then divided into frames with a duration of eight hundred milliseconds using a Hann window with seventy-five percent overlap between adjacent frames. The first thirty MFCCs are retained for each frame and ΔMFCCs are calculated by finding the difference between the coefficients for adjacent frames.

The frames without any vocal activity are identified and removed so as not to bias the features with periods of silence or instrumental only sections. This can be achieved by using either the "full-reference" method or the "no-reference" method. The full-reference method uses the unmixed vocals track to detect frames with presence of acoustic activity by means of thresholding the energy present in each of the analysis-frames. This serves as a crude but effective acoustic activity detector which allows for the frames without any vocal activity to be removed.

The no-reference method uses a classifier designed to detect regions without vocal activity using the method described by Lehner et al. [40] and operates on the original musical mix. This method uses the MFCC based features similar to the ones described above; with only difference is that instead of the extracted vocals, the features are extracted from the original audio mix. In this method a support vector machine was trained using these features which classifies the windowed frames into vocal and non-vocal frames.

Unlike Lehner et al. [40] who use a support vector machine based classifier, a

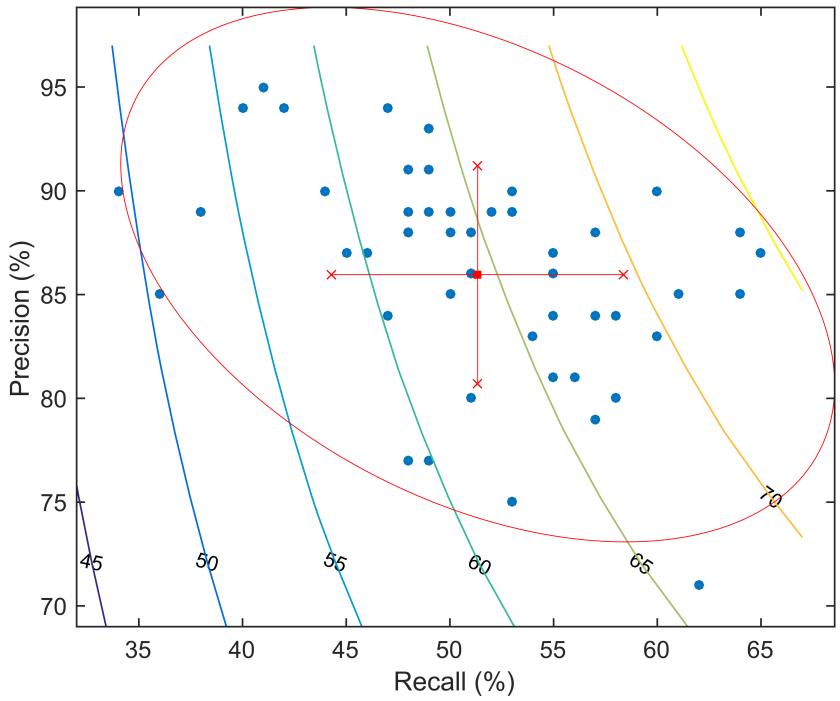


Figure 6.1: Precision v. Recall for fifty iterations of vocal v. non-vocal classifier. The cross shows the mean and standard deviation, ellipse is the 95% confidence bound and contours are the F1-statistic.

random-forest classifier with one-hundred and twenty trees was used in this implementation as it was found to perform better. This classifier was trained using one-hundred songs from a mixed set of samples from MedleyDB Multitrack Dataset [3] and iKala Dataset [7]. As excerpts from MedleyDB Dataset are also used as a part of the listening tests described in Chapter 3, it was ensured that the two sets did not overlap in song and artist choices. The validation statistics of fifty iterations of randomly partitioned songs for training vs. testing sets (with twenty-five percent holdout for testing) are shown in Figure 6.1.

The figure shows a scatter of precision and recall (with non-vocal as positive class) for all the fifty iterations, with the average value denoted by the intersection point of the cross and standard deviation of precision and recall denoted by the arms of the cross. The ninety-five percent confidence bound (assuming multivariate normal distribution), shown as the ellipse, indicates that the worst case performance of the

classifier results in precision values of about seventy-three percent. The classification scheme was biased such that the cost of incorrectly classifying non-vocal frames as vocals was twice that of misclassifying the vocal frames. This is important because when the vocal frames are used for calculating VIS, the inclusion of non-vocal frames in the set degrades the result much more than missing a few frames which do contain vocals.

After the non-vocal frames have been removed, the means and standard deviations of the first thirty MFCCs and ΔMFCCs (including zero order MFCC) are found. This results in a feature vector of one-hundred and twenty dimensions for each of the extracted vocal excerpts. These features will be referred to as MFCC_μ^k, MFCC_σ^k, ΔMFCC_μ^k and, $\Delta\text{MFCC}_\sigma^k$ (where k is the order of the coefficient from 0 to 29) in the remainder of this section. No statistically significant differences were found between the features from the "full-reference" and the "no-reference" methods using paired t-tests.

6.1.2 Feature Sensitivity to Instrumental Mixing Levels

To be useful for VIS, it is important that the above features are sensitive to the level of instrumental presence in the extracted vocals. To test this, audio clips, from twenty-five songs, were mixed with different levels of instrumentals as compared to vocals at +5 dB, 0 dB, and -5 dB. These clips were normalized such that the overall loudness of each of the clips was the same. The MFCC based features extracted from all these clips, along with the vocals only clips ($-\infty$ dB) and the instrumentals only clips ($+\infty$ dB), were compared using pair-wise Kolmogorov-Smirnov tests ordered by increasing instrumental loudness that are shown in Table 6.1. The comparisons for which the null hypothesis (that the two sets of features belong to identical distributions) can be rejected with $\alpha = 0.05$ are marked with "x".

Table 6.1 shows that differences for MFCCs and ΔMFCCs are not significant

Table 6.1: Results from paired Kolmogorov-Smirnov test comparing the four categories of features for decreasing level of relative instrumental loudness. Statistically significant values with $\alpha = 0.05$ marked as x.

Instr. Levels	MFCC_μ	MFCC_σ	ΔMFCC_μ	ΔMFCC_σ
$+\infty$ dB v. +5 dB	-	x	-	x
+5 dB v. 0 dB	-	x	-	x
0 dB v. -5 dB	-	x	-	x
-5 dB v. $-\infty$ dB	x	x	x	x

when comparing between clips with different levels of instrumentals, and only show significant differences between the feature distributions when comparison is between the vocals-only clips and clips with instrumentals mixed in. On the other hand, the standard deviations of MFCCs and ΔMFCCs show statistically significant variations with different levels of instrumentals present in the audio clips, and therefore are sensitive to the relative loudness of the instrumentals present in the clips.

This analysis shows that both the mean and standard deviation based features are useful as predictors for VIS. While the means of MFCCs and ΔMFCCs help detect the presence or absence of residual instrumental music, the standard deviations provide indicators for the loudness level of the residual audio.

6.1.3 Calculating VIS using Regression

To calculate VIS, the features described in the previous section are extracted from the audio excerpts used in the listening experiments. A regression model is trained using these features, with the ratings from the vocal isolation task (T2) as the target, to get prediction score. The prediction score obtained from the regression model is VIS.

It was observed that for both the training and the testing data, the one-hundred and twenty dimensional features are highly redundant, with high inter-feature covariance. In order to use these features for regression, it is advisable for the feature set to be reduced to a smaller dimensionality such that the redundant information

is removed [27]. For this purpose, principal component analysis (PCA) is performed on the features matrix from the training data (from LE1) [35]. Applying PCA to the original set of features, results in a new set of features each of which is a linear combination of the original features. The new features generated using PCA have the property that most of the information in the data can be explained by the first few features and the remaining features are redundant. In the case of the LE1, it was determined that the first fifteen features explained around ninety percent of the underlying variance and were used to train the regression model. Adding extra features did not improve the model performance with the performance degrading if the dimensionality exceeded eighteen features.

VIS is calculated as the result of least-squares regression with the first fifteen features from the PCA features as the predictors, where is is the predicted value obtained from the regression model. The target variable for the regression model is the set of representative subjective ratings for each of the audio excerpts obtained from the listening experiments as described in Section 4.2 (Chapter 4).

6.2 Performance Evaluation for VIS

For calculating VIS, the features as described in Section 6.1.1 were extracted from the audio excerpts used in the listening experiments LE1, LE2 and LE3. The regression models of VIS were trained using the results of LE1 and tested against the results from LE2, and the repeatability of the results was validated using LE3.

The same evaluation methodology that was used to assess the performance of the current state-of-the art objective measures (BSS_Eval and PEASS) in Chapter 5 was used to assess the performance of VIS. As the results from LE1 were used for training the regression model for VIS, evaluation in terms of Pearson's correlation coefficient (PCC) and Spearman's correlation coefficient (SCC) was performed with the ratings from LE2 and LE3.

Table 6.2: Correlation between VIS and vocal isolation (T2) ratings. Results for SIR from BSS_Eval measures and OPS from PEASS measures are reproduced from Chapter 5 for comparison.

		VIS	**SIR**	**IPS**
Pearson's Correlation (PCC)	**LE2**	0.736 [-.12,+1.0]	0.614 [-.08,+.98]	0.633 [-.20,+1.0]
	LE3	0.718 [-.08,+1.0]	0.444 [-.62,+1.0]	0.386 [-.80,+.98]
Spearman's Correlation (SCC)	**LE2**	0.664 [0.00,+1.00]	0.614 [-.08,+.98]	0.638 [-.22,+1.0]
	LE3	0.621 [+.07,+1.0]	0.383 [-.63,+.99]	0.379 [-.71,+1.0]

Table 6.2 shows the result of comparing the objective VIS score with vocal isolation (T2) from LE2 and LE3. The results shown in the table use the "no-reference" method for evaluation of VIS. The table also reproduces from Chapter 5 the similar assessment for the current state-of-the objective measure to for direct comparison. The performance of VIS is compared against the performances of SIR (which is a part of BSS_Eval measures) and IPS (from PEASS measures) both of which are designed to measure interference in separated signal (vocals) from other sources (accompaniment).

The results show that the average values for both PCC and SCC are higher for VIS than for either of the state-of-the-art measures for estimating interference. Although VIS does not demonstrate a statistically significant Pearson's correlation (as described in Chapter 5) against the subjective vocal isolation ratings, the PCC confidence intervals for VIS are narrower than the corresponding intervals for either SIR or IPS. This demonstrates that for the purpose of evaluating vocal isolation in singing voice separation VIS performs better than the current state-of-the-art objective measures. In contrast to current state-of-the-art source separation measures, the computation of VIS does not require a reference (i.e., unmixed vocal and/or instrumental track),

which makes it more suitable for real-world applications.

CHAPTER VII

DEVELOPMENT OF INTELLIGIBILITY MEASURE

It is an important requirement for any SVS process that the intelligibility of the extracted vocals is not degraded by the source separation process. Although neither the BSS_Eval measures or the PEASS measures, the two current state-of-the-art sets of objective measures used for evaluating SVS implementation, directly evaluate the preservation of intelligibility in the separated vocals, it is maintained that vocal intelligibility is perceptually important in itself even if it is not a measure of quality of separation [64].

This chapter addresses the issue by introducing a new objective measure, Vocal Intelligibility-Preservation Score (VIPS), to evaluate the performance of various SVS implementations in context of preservation of lyric-intelligibility in separated vocals. It should be noted that the assessment of perceived preservation of intelligibility in vocal extraction using SVS is different than traditional speech intelligibility assessment techniques. Traditional techniques such as Modified Rhyme Test (MRT) [23], Diagnostic Rhyme Test (DRT) [65], or Speech Reception Threshold (SRT) [48] are not applicable in this case as these techniques depend upon the listener having no *a priori* knowledge of the speech content, a condition that is easily violated in the case of multiple comparisons. Here, the subjective evaluation is aiming to compare the degree of preservation of perceived intelligibility in the separated vocals when compared to the original as a reference.

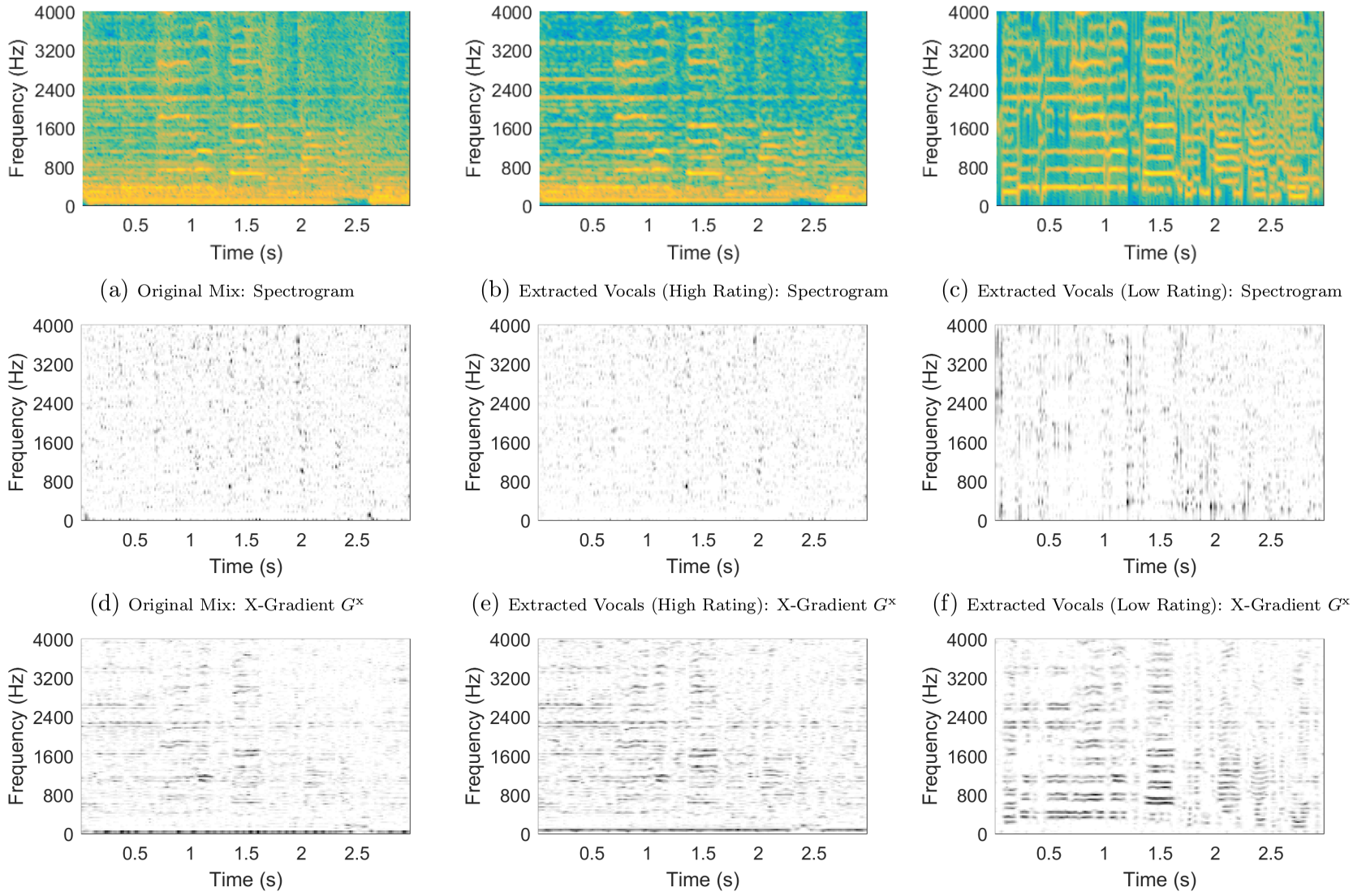
VIPS has been designed as a "no-reference" measure, i.e., it does not need the unmixed vocal and instrumental tracks for the evaluation. The development process for VIPS is detailed in Section 7.1 which starts the discussion by explaining the

motivation behind the design of VIPS and continues on to describing the process of feature extraction Section 7.1.1. Following the feature extraction process, the regression model used for combining the features is described in Section 7.1.2. The latter part of the chapter discusses evaluation of the new measure in Section 7.2 and explores how its performance compares to the existing state-of-the-art objective measures.

7.1 Design Process

Section 4.2 in Chapter 4 analyzed how the different subjective assessment tasks used for the listening experiments (see Section 3.1.3) compare to each other. It was determined from this analysis that one of the major factors contributing to the perceived loss of intelligibility in the extracted vocals was the presence of artifacts (additive processing noise) in the output of the SVS algorithms. As explained in Section 2.1 of Chapter 2, such artifacts commonly take the form of abrupt glitch like noises in the audio signal, also known as "musical noise". This is due to presence of isolated peaks in the spectrum of the signal and is usually found in outputs of processes involving spectral subtraction or spectral masking [60]. This kind of time-variant spectral processing involving methods (such as Harmonic-Percussive Source Separation (HPSS) [34], constrained Non-negative Matrix Factorization (NNMF) [46], etc.) is very common in SVS algorithms which use these techniques to isolate the contributions of the sources of interest [14].

Figures 7.1(a-c) show examples of the short-time Fourier transform (STFT) spectrogram from the same excerpt for the mixed signal and two extracted vocals using two different systems. For this excerpt the intelligibility of the first result (Fig. 7.1b) is rated higher than the second (Fig. 7.1c). It is observed that there is an increase in abruptness in the spectral content as intelligibility decreases, i.e., there is a sharper contrast in the spectral content of the less intelligible samples as compared to the



(a) Original Mix: Spectrogram

(b) Extracted Vocals (High Rating): Spectrogram

(c) Extracted Vocals (Low Rating): Spectrogram

(d) Original Mix: X-Gradient G^{x}

(e) Extracted Vocals (High Rating): X-Gradient G^{x}

(f) Extracted Vocals (Low Rating): X-Gradient G^{x}

(g) Original Mix: Y-Gradient G^{y}

(h) Extracted Vocals (High Rating): Y-Gradient G^{y}

(i) Extracted Vocals (Low Rating): Y-Gradient G^{y}

Figure 7.1: Comparison of short time spectrogram of the original mix and extracted vocals from two SVS implementations with different intelligibility and the results of convolving each with the two Sobel gradient operators g^{x} and g^{y}

more intelligible ones.

These observations are the motivation behind the design of the proposed evaluation measure VIPS. The following section discusses some features that have been developed to capture the effects of these observations.

7.1.1 Feature Extraction

As stated above, the principal cause of degradation in intelligibility in the extracted vocals compared to the original audio is the presence of artifacts. Such artifacts make themselves visible in short-time spectrograms as abrupt discontinuities along both time and frequency axes. This is reminiscent of edges in images, and is the inspiration behind using standard edge detection algorithms prevalent in image processing literature to detect these abrupt spectral artifacts. Various image processing algorithms like Sobel's algorithm [18, 54], Canny's algorithm [4], etc. were tried; it was determined experimentally that the features obtained from Sobel's algorithm were the most correlated with the results from the intelligibility assessment task (T5) of LE1. The extraction process for these features is described below.

For the extraction of the features, the extracted vocal signal $\hat{x}_{\mathrm{v}}$ and the mixed signal x_{m} are passed through an anti-aliasing filter with bandwidth of 8 KHz and are re-sampled to 16 KHz (if the sampling rate is different). This process removes any high frequency noise above 8 KHz from the signal. Since the vocals along with the audible harmonics rarely exceed 8 KHz in frequency content, the artifact-noise above this frequency is expected to have little to no affect on the perceived intelligibility.

For both the signals $\hat{x}_{\mathrm{v}}$ and x_{m}, the short-time Fourier transforms $\hat{X}_{\mathrm{v}}(n,k)$ and $X_{\mathrm{m}}(n,k)$ are computed by using thirty-two milliseconds Hamming window with a fifty percent overlap for the n^{th} DFT bin and the k^{th} analysis window. These parameters provide an adequate resolution in both time and frequency domains such that the musical noise is not smeared out in either dimension, which might be the case if

window lengths used are too long or short. For each STFT, the power spectrogram is calculated as shown in Eqn. 7.1 below.

$$P_{\mathrm{j}}(n,k) = 20\log_{10}|X_{\mathrm{j}}(n,k)| \tag{7.1}$$

Two Sobel operators g^{x} and g^{y}, as defined below [18, 54], are used in order to determine the edges in the power spectrogram.

$$g^{\mathrm{x}} := \frac{1}{8}\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{7.2a}$$

$$g^{\mathrm{y}} := \frac{1}{8}\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{7.2b}$$

The x and y gradients of the power spectrograms are computed by their two-dimensional convolution with each of the Sobel operators as shown in Eqn. (7.3a) 7.3d. The gradient $G_{\mathrm{v}}^{\mathrm{x}}$ highlights the sharp discontinuities perpendicular to the time axis in the power spectrogram and the gradient $G_{\mathrm{v}}^{\mathrm{y}}$ does the same for discontinuities perpendicular to the frequency axis. This is shown in Figs. 7.1(d-f) and 7.1(g-i), respectively. $G_{\mathrm{m}}^{\mathrm{x}}$ and $G_{\mathrm{v}}^{\mathrm{m}}$ perform the same functions for the mix signal, x_{m}.

$$G_{\mathrm{v}}^{\mathrm{x}}(n,k) = \hat{P}_{\mathrm{v}}(n,k) * g^{\mathrm{x}} \tag{7.3a}$$

$$G_{\mathrm{m}}^{\mathrm{x}}(n,k) = P_{\mathrm{m}}(n,k) * g^{\mathrm{x}} \tag{7.3b}$$

$$G_{\mathrm{v}}^{\mathrm{y}}(n,k) = \hat{P}_{\mathrm{v}}(n,k) * g^{\mathrm{y}} \tag{7.3c}$$

$$G_{\mathrm{m}}^{\mathrm{y}}(n,k) = P_{\mathrm{m}}(n,k) * g^{\mathrm{y}} \tag{7.3d}$$

As the human auditory system perceives different loudness for equally intense audio signals at different frequencies, the edges can be weighted perceptually. A

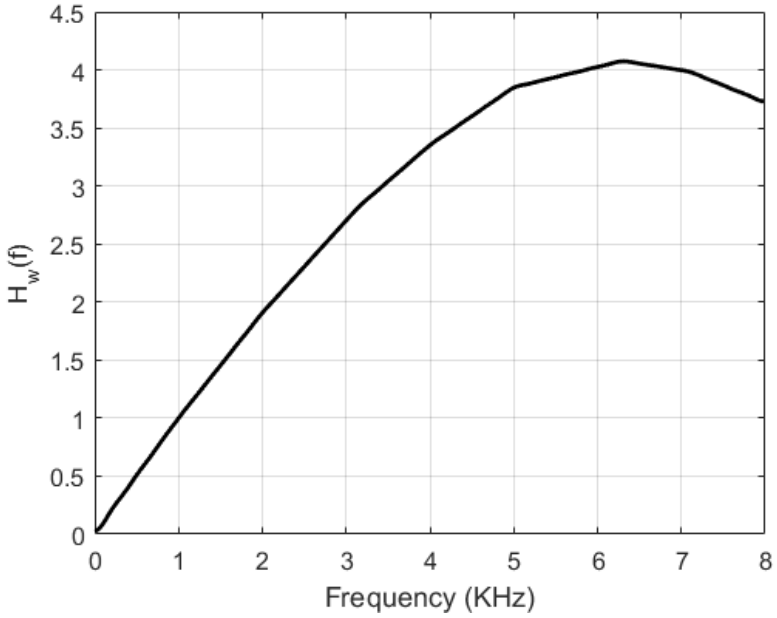


Figure 7.2: The magnitude response of the perceptual loudness weighting filter based on ITU-R BS.468-4

weighted loudness FIR filter is designed using recommendation ITU-R BS.468-4 [33] is used for this purpose. Its magnitude frequency response is defined as $H_{\mathrm{w}}(f)$ or the discrete version $H_{\mathrm{w}}[n]$ which is computed for f equals f_{n}, the central frequency of the n^{th} DFT bin. This is shown in Fig. 7.2. Alternatively, the edges can be weighted uniformly (as opposed to perceptual weighting) with weights of one. In this case $H_{\mathrm{w}}[n]$ is one for all values of n.

Edges in the spectrum are determined by thresholding the gradients obtained in equations 7.3a to 7.3d in two ways. First, the threshold values $\mathcal{T}_{\mathrm{x}}$ and $\mathcal{T}_{\mathrm{y}}$ are defined and the edge position matrices are calculated for both the separated vocal and the mix signal in x and y directions as shown in Eqs. 7.4a and 7.4b, where κ_{x} and κ_{y} are constants between zero and one. The results in this implementation were obtained using $\mathcal{T}_{\mathrm{x}}$, $\mathcal{T}_{\mathrm{y}}$, κ_{x} and, κ_{y} as 10, 5, 0.8 and, 0.6 respectively, with their values determined experimentally.

$$E_{\mathrm{x}}(n,k) = \begin{cases} H_{\mathrm{w}}[n] & \text{if } G^{\mathrm{x}}_{\mathrm{v}}(n,k) \geq \mathcal{T}_{\mathrm{x}} \quad \& \\ & \quad G^{\mathrm{x}}_{\mathrm{m}}(n,k) < \kappa_{\mathrm{x}}\mathcal{T}_{\mathrm{x}} \\ 0 & \text{otherwise} \end{cases} \tag{7.4a}$$

$$E_{\mathrm{y}}(n,k) = \begin{cases} H_{\mathrm{w}}[n] & \text{if } G^{\mathrm{y}}_{\mathrm{v}}(n,k) \geq \mathcal{T}_{\mathrm{y}} \quad \& \\ & \quad G^{\mathrm{y}}_{\mathrm{m}}(n,k) < \kappa_{\mathrm{y}}\mathcal{T}_{\mathrm{y}} \\ 0 & \text{otherwise} \end{cases} \tag{7.4b}$$

Since the purpose of detecting these edges is to obtain features which are perceptually relevant to loss of intelligibility, it is advisable to remove effect of edges that may not be audible to the listener by the virtue of being masked by other sounds present in the signal. This optional step is performed before weighting the edges (perceptually or uniformly). A simultaneous global masking threshold for the separated vocals $M(n,k)$ is calculated along with the loudness estimate $L(n,k)$ (in dbSPL) for each STFT frame for $\hat{x}_{\mathrm{v}}$ using the MPEG I standard psychoacoustic model II specification[1] [31]. The edges $E_{\mathrm{x}}(n,k)$ or $E_{\mathrm{y}}(n,k)$ with a loudness estimate below the global masking threshold are disregarded by equating them to zero.

The following features are extracted from E_{x} and E_{y} as averages of the edge position matrices as given in Eqs. 7.5a and 7.5b, their standard deviation across time as given in Eqs. 7.5c and 7.5d, and across frequency components, Eqs. 7.5e and 7.5f. Here, N is the number of DFT bins and K is the number of analysis frames.

$$\mathrm{AVX} = \frac{1}{KN}\sum_{k=0}^{K-1}\sum_{n=0}^{N-1} E_{\mathrm{x}}(n,k) \tag{7.5a}$$

$$\mathrm{AVY} = \frac{1}{KN}\sum_{k=0}^{K-1}\sum_{n=0}^{N-1} E_{\mathrm{y}}(n,k) \tag{7.5b}$$

[1] Jon Boley's implementation was used (`http://tiny.cc/jboley`)

$$\mathrm{STX} = \sqrt{\frac{1}{K-1}\sum_{k=0}^{K-1}\left(\frac{1}{N}\sum_{n=0}^{N-1}E_{\mathrm{x}}(n,k) - \mathrm{AVX}\right)^2} \tag{7.5c}$$

$$\mathrm{STY} = \sqrt{\frac{1}{K-1}\sum_{k=0}^{K-1}\left(\frac{1}{N}\sum_{n=0}^{N-1}E_{\mathrm{y}}(n,k) - \mathrm{AVY}\right)^2} \tag{7.5d}$$

$$\mathrm{SFX} = \sqrt{\frac{1}{N-1}\sum_{n=0}^{N-1}\left(\frac{1}{K}\sum_{k=0}^{K-1}E_{\mathrm{x}}(n,k) - \mathrm{AVX}\right)^2} \tag{7.5e}$$

$$\mathrm{SFY} = \sqrt{\frac{1}{N-1}\sum_{n=0}^{N-1}\left(\frac{1}{K}\sum_{k=0}^{K-1}E_{\mathrm{y}}(n,k) - \mathrm{AVY}\right)^2} \tag{7.5f}$$

7.1.2 Regression Modeling

Linear regression is used to calculate VIPS from the features described above. The normalized representative scores, from the preservation of intelligibility task (T5) for the listening experiments, are used as the target for the regression with the features AVX, AVY, STX, STY, SFX and SFY as predictors. The process of obtaining the normalized representative scored from the subjective ratings is described in Section 4.2.

From this process VIPS is obtained in the form of a linear combination of the features as shown in Eqn. (7.6).

$$\mathrm{VIPS} = a_0 + \sum_{i=1}^{6} a_i F_i \tag{7.6}$$

Here, a_0 to a_6 are constants found by the regression process and F_1 to F_6 are the feature values for AVX, AVY, etc.

7.2 Performance Evaluation for VIPS

The evaluation methodology for VIPS follows a similar procedure as described for Vocal Isolation Score (VIS) in Chapter 6. The performance for VIPS is assessed in terms of the distribution of its correlation with the subjective ratings in terms of Pearson's correlation coefficient (PCC) and Spearman's correlation coefficient (SCC). The exact methodology is detailed in Chapter 5.

Table 7.1: Comparison between different options for calculating VIPS. The option with the highest correlation is highlighted using bold typeface.

Option #	**Loudness Weighting**	**Masking Threshold**	**Pearson's Correlation**		**Spearman's Correlation**	
			LE2	**LE3**	**LE2**	**LE3**
1	Perceptual	Yes	0.66 [0.19,1.00]	0.61 [-0.06,1.00]	0.682 [0.09,1.00]	0.613 [-0.08,1.00]
2	Perceptual	No	0.841 [0.56,1.00]	0.398 [-0.64,0.94]	0.796 [0.28,1.00]	0.305 [-0.78,0.96]
3	**Uniform**	**Yes**	**0.825 [0.42,1.00]**	**0.818 [0.43,1.00]**	**0.769 [0.36,1.00]**	**0.778 [0.29,1.00]**
4	Uniform	No	0.873 [0.59,1.00]	0.759 [0.36,1.00]	0.813 [0.36,1.00]	0.672 [0.16,1.00]

The features for VIPS are extracted for the audio excerpts used in the three listening experiments, and the regression model is trained using the results from the preservation of intelligibility task (T5) of LE1. The ratings from LE2 and LE3 are used to test the model and validate the final result respectively. Since there are multiple optional steps in the procedure used for calculating VIPS, multiple regression models are trained with and without perceptual weighting, and also with and without psychoacoustic thresholding. The results of training the models using features from LE1 in each case and testing on LE2 and LE3 are shown in Table 7.1.

The results in the table show that extracting the features using different methods does indeed affect the correlating between the subjective ratings and VIPS. For option two in the table, where the edges were perceptually weighted but the masking threshold wasn't applied, the results between the test set (LE2) and the validation set (LE3) do not match. For option one, with both perceptual weighting and auditory masking used, the performance matches for LE2 and LE3 but fails to provide statistically significant correlation (see Chapter 5) in the case of LE3. The results calculated using either of options three or four demonstrates that VIPS is highly correlated with the perceptual ratings of intelligibility, and that the correlation for both of them is statistically significant. Although the difference between the performance of VIPS

Table 7.2: Correlation between VIPS and vocal intelligibility (T5) ratings. Results for SAR from BSS_Eval measures and APS from PEASS measures are reproduced from Chapter 5 for comparison.

		VIPS	**SAR**	**APS**
Pearson's Correlation (PCC)	**LE2**	0.825 [+.42,+1.0]	0.466 [-.40,+.99]	0.593 [-0.10,+1.00]
	LE3	0.818 [+.43,+1.0]	0.439 [-.52,+.97]	0.612 [-0.13,+1.00]
Spearman's Correlation (SCC)	**LE2**	0.769 [+.36,+1.0]	0.466 [-.40,+.99]	0.601 [-0.15,+1.00]
	LE3	0.778 [+.29,+1.0]	0.403 [-.59,+.99]	0.560 [-0.24,+1.00]

calculated using options three and four is not statistically significant, option three is chosen for further comparison with the current state-of-the-art measures as the best performing candidate.

Table 7.2 shows the comparison among the performance of VIPS, Source to Artifact Ratio (SAR) (see Section 2.2.1) and Artifacts-related Perceptual Score (see Section 2.2.2). Although, neither SAR or APS are objective measures that are designed to measure the preservation of intelligibility, they are used here for comparison because they characterize the performance of source separation algorithms in terms of presence of artifacts. As the intelligibility of the extracted vocals is highly correlated to the presence of artifacts, SAR and APS have been used due to the lack of objective measures which provide direct comparison.

It is seen from the results in Table 7.2 that VIPS easily outperforms both SAR and APS when it comes to evaluating the performance of SVS algorithms in context of preservation of intelligibility of the vocals. VIPS has higher average PCC and SCC values for both LE2 and LE3 as compared to either SAR or APS. Additionally, the confidence intervals for VIPS in all the cases are all in the positive region indicating

that VIPS performs consistently for predicting the perceived intelligibility for SVS algorithms.

The greatest advantage VIPS has over the current state-of-the-art methods is that it can be calculated in the absence of any reference signals in the form of unmixed vocals or instrumentals. This immensely increases the value of VIPS as a measure of performance for SVS techniques in real world applications where the unmixed audio is generally not available.

CHAPTER VIII

CONCLUSION

8.1 Research Summary

Singing Voice Separation (SVS) uses audio source separation methods to isolate the vocal component from the background accompaniment in a song mix. A key challenge currently associated with evaluation of SVS is a lack of objective measures which correlate consistently with subjective evaluation. Additionally, the current state-of-the-art evaluation measures require the use of unmixed vocal and instrumental tracks which are often not available. The research presented in this thesis is an attempt to address these challenges by introducing two new objective measures for evaluation of SVS without requiring the use of reference audio tracks containing the unmixed vocal or instrumental music.

A preliminary listening experiment (LE1) was designed to provide a listener based subjective assessment of performance of various SVS algorithms, and to analyze how the state-of-the-art objective measures compared to human judgment. For the listening experiment audio excerpts from pop-music songs were processed with different SVS algorithms and the participants were asked to judge the performance of these extracted vocal clips and compare them to each other.

Although a statistically significant correlation (0.74 for Pearson's and 0.81 for Spearman's correlation) was found between SAR of BSS_Eval measures [14] with the ratings from the LE1 task designed to assess the preservation of intelligibility in the extracted vocals after SVS; neither SIR nor NSDR, however, demonstrated a significant correlation with the subjective assessment of vocal isolation or the overall quality of separation. Similarly, in the case of PEASS measures [14, 63], while APS

correlated strongly with the preservation of vocal intelligibility (0.56 for Pearson's and 0.80 for Spearman's correlation), the other measures OPS and IPS failed to show statistically significant correlations with the other two tasks.

Based on the results of LE1, it was determined that there existed a need for developing new objective quality assessment measures specially designed to evaluate the performance of SVS algorithms. Additionally, from listening to the separated vocals as well as feedback from the participants in LE1, it was determined that the impairments produced by the SVS algorithms are very severe. Although a reference signal is generally necessary for nuanced comparison between two similar audio clips, with the severe degradation produced by SVS algorithms it was deemed worth exploring if the quality in this case could be assessed without requiring a reference.

Two new objective measures were introduced as a part of this research. The Vocal Isolation Score (VIS) was designed to assess the quality of isolation produced by various SVS algorithms when separating the vocals from the accompaniment. VIS was constructed using MFCC based features to train a regression model with the ratings from the vocal isolation task from LE1 as the target variable. Similarly, Vocal Intelligibility Preservation Score (VIPS) was developed to evaluate the amount of intelligibility of the vocals preserved during the SVS process. VIPS was designed to exploit the fact that most of the loss in intelligibility in the separated vocals is due to the presence of spectral processing artifacts. The effect of these spectral artifacts was captured by using features inspired from edge detection algorithms used in image processing, and were used to build a regression model to calculate VIPS with the scores obtained from the preservation of intelligibility task in LE1 as the target. VIPS fills in a gap existing in the community by providing an objective evaluation of intelligibility in context of SVS as none of the current state-of-the-art objective measures are designed for this task.

In order to compare the performance of VIS and VIPS in comparison to BSS_Eval and PEASS measures, two new listening experiments (LE2 and LE3) were conducted. The purpose behind these tests was to provide testing and validation data for the regression models trained for VIS and VIPS. To this end it was ensured that none of the excerpts that were used for LE1 were used again for these two. The two new experiments also made use of three SVS algorithms which were not used for LE1. This was done to test if the models for VIS and VIPS that were trained using results from LE1 provided robust results when used with new and unseen data.

It was found upon comparison between the results of evaluating vocal isolation using VIS, SIR, and IPS (all of which are designed to assess the interference from other sources in the separated signal) that VIS outperformed both SIR and IPS. VIS was shown to have average Pearson's and Spearman's correlation of 0.74 and 0.66 respectively with results from LE2, and 0.72 and 0.62 for LE3. Neither SIR nor IPS showed similarly high correlations with the ratings from LE2 or LE3 as detailed in Section 6.2.

While there were no objective measures available that provided a direct evaluation of intelligibility, the performance of VIPS was compared to SAR and APS. This is due to the fact that loss of intelligibility was shown to be highly correlated with the presence of artifacts as shown in Section 4.2. Additionally, both SAR and APS have shown to be significantly correlated with the results of the preservation of intelligibility task from LE1. Upon comparing VIPS, SIR and APS against the results of the intelligibility task from LE2 and LE3, VIPS was found once again to be the better performer and demonstrated a higher correlation with the subjective ratings than SIR or APS. VIPS was shown to have statistically significant correlation with the subjective ratings. The average Pearson's correlation coefficient between VIPS and the subjective ratings was 0.825 and 0.818 for LE2 and LE3 respectively with corresponding Spearman's correlation averages being 0.769 and 0.778. Neither SAR

nor APS performed as well with having average correlations in the range of 0.4 to 0.7 as detailed in Section 7.2.

From this thesis it can be concluded that the two new measures introduced here, VIS and VIPS, fill the gap that existed in the signal processing community regarding the perceptual evaluation of singing voice separation. Other than an improvement upon the state-of-the-art in terms of performance both VIS and VIPS have the additional advantage that they do not require references in the form of unmixed vocal or instrumental tracks to perform objective evaluation which is truly a novelty among all the objective measures used for evaluating source separation.

8.2 Contributions and Future Work

This thesis set out to address the challenge of not having objective measures for evaluation of SVS which are consistent with human assessment of quality of the extracted vocals. Both the objective measures introduced here were demonstrated to be an improvement over the current state-of-the-art, with VIPS providing extremely consistent way (in terms of statistically significant correlation) of predicting the quality of SVS implementations in context of preserving the intelligibility in the separated vocals.

Both VIS and VIPS were able to improve the state-of-the-art by exploiting prior domain knowledge associated with SVS. In the case of VIS, the fact that MFCC based features have significant differences between them in the presence or absence of instrumental music was used to formulate a new measure which was able to measure the amount of interference present in the separated vocals by reacting to the presence of residual instrumental presence. VIPS exploited the fact that loss of intelligibility in the separated vocals was highly correlated with the amount of spectral artifacts (known as "Musical Noise") introduced by time-frequency masking in the SVS process. In contrast to this, both BSS_Eval and PEASS measures do not assume

any prior knowledge about the signals being separated or the process of separation. While this approach may result in more generalized uses of these measures, it was demonstrated by the research presented that using domain specific knowledge can result in better performance for specific use cases.

Another way the research presented here contributes to the community is by serving as a first proof of concept demonstrating that the use of reference audio is not necessary for objective measures for source separation dealing with severe impairments. This is important because most real-world applications of SVS or other source separation tasks are demanded in a scenario where the unmixed reference audio is not available. To evaluate the performance in such cases necessitates the use of no-reference objective measures.

Although the new measures introduced in this thesis have addressed the challenges of evaluating SVS in terms of isolation and intelligibility, work needs to be done in the future to develop new measures which provide perceptually relevant assessment of SVS in terms of overall quality as well as additive and subtractive distortions. Another area where further research is required is the evaluation of the separated accompaniment, which is the residual signal remaining after the vocals have been separated. The data from the three listening experiments will be made available for future research and investigation regarding objective and subjective evaluation of singing voice separation and its associated fields.

APPENDIX A

DETAILS OF LISTENING EXPERIMENTS

A.1 Listening Experiment I (LE1)

A.1.1 Audio Data

Excerpts of five to ten seconds duration from eight songs from the Medley DB multitrack dataset were used [3]. The excerpts were manually chosen to ensure that they contained both vocal and instrumental portions. Four of the excerpts were from "Singer/Song Writer" genre and the other four were from "Rock" genre.

A.1.2 SVS Algorithms

Four SVS algorithms were used to process each of the excerpts. These algorithms were chosen because of their superior performance in MIREX '14 competition, SVS task [10] The algorithms that were used in LE1 are Ikemiya et al. [29], Jeong and Lee [34], Rao et al. [52], and Rafii and Pardo [50]. All the excerpts were processed by either author provided implementations of the algorithms or the authors themselves.

A.1.3 Participant Details

Subjects were gathered from a normal hearing population of graduate and undergraduate students, with ages varying from nineteen to thirty-six, to participate in the experiment. Out of thirty participants, eleven had experience in a music related field and six were professionally trained in music and/or had studio recording experience. The others were not trained in music. The number of male participants was twenty-five, while five were female.

A.1.4 MUSHRA Interface

The interface used for conducting the experiment is shown in Figure A.2.

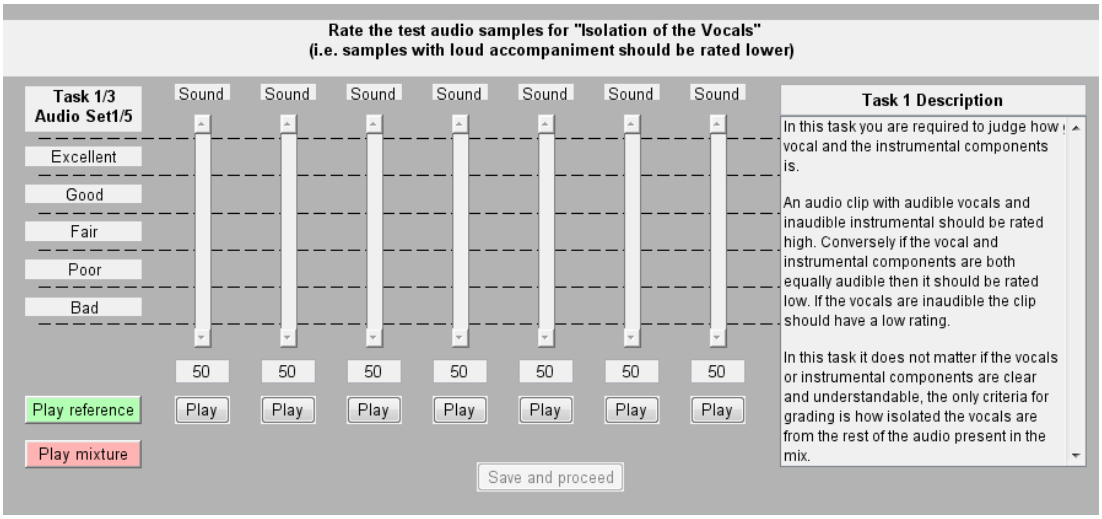


Figure A.1: The MUSHRA interface used in the first listening experiment (LE1).

A.1.5 Task Descriptions

The listening test comprised of three tasks which were, in order, vocal isolation task (T2), vocal intelligibility task (T5) and overall separation quality task (T1). The overall quality task was presented last to the participants because it was expected that this would result in more representative subjective ratings as the participants would be forced to consider multiple factors when assessing the audio clips. Due to lack of agreement among the raters for task T1 (as discussed in Section 4.1), the order was switched in the latter experiments.

Each participant was presented with a short video before starting the testing session, where the tasks and their requirements were described. Additionally, the participants were provided with a written copy of the instructions for each of the tasks on the screen as a part of the interface.

A.2 Listening Experiment II (LE2)

A.2.1 Audio Data

In this case the audio data consisted of six excerpts from the Medley DB dataset [3] and six from the "The Open Multitrack Testbed (OMT)" [9]. Each excerpts was five to ten seconds long and contained overlapping vocal and instrumental music. It was ensured that the songs used in LE1 from MedleyDB dataset were not repeated in LE2. All the songs from both the datasets were restricted to "Singer/Song Writer" or "Pop Music" genres.

A.2.2 SVS Algorithms

The number of SVS algorithms used for processing was increased to five in LE2. Two of the SVS algorithms were from the earlier set used for LE1, i.e., Jeong and Lee [34], and Rafii and Pardo [50]. The three new algorithms chosen for separating the vocals were Huang et al. [25], McVicar et al. [43], and Liutkus et al. [42].

A.2.3 Participant Details

The listening experiment was published on the Internet and publicized on social media, forums, and mailing lists. A total of one hundred and thirty-two submissions were received of which, only ninety-one were complete submissions. The age for the participants varied from nineteen to sixty-nine, with majority in the age range of twenty-two to thirty-four. The average time for completing the listening experiment which constitutes grading two excerpts across five tasks was fifteen. Of the ninety one participants who completed the experiment, twenty four reported themselves to be experienced at recording/producing music and forty participants reported some familiarity with vocal/instrumental music.

A.2.4 MUSHRA Interface

The interface used for conducting the experiment is shown in Figure A.2. The interface was implemented for use inside web browsers and was modified for the purpose of the listening experiments from its original source [38]. It was ensured that the interface remained responsive and usable, irrespective of the web-browser or viewing device used. Participation was reported across various devices such as computers, mobile phones, and touchscreen tablets.

A.2.5 Task Descriptions

All five tasks (as described in Chapter 3) were a part of this experiment. In this case the overall quality of separation task (T1) was presented first to the participants. This was to gather an unbiased opinion of the subjects about the subjective quality of vocal separation, without being influenced by judgment across different categories. Although, moving the order of the tasks did improve the median for Krippendorff's Alpha (KA) values for LE2 over LE1 in T1, the ratings still did not show enough improvement to be considered .

Prior to starting the listening experiment, a brief overview of each of the tasks was provided to the participants. Detailed instructions for each of the tasks were made available as a part of the on-screen user interface during the testing process.

A.3 Listening Experiment III (LE3)

A.3.1 Audio Data

The audio excerpts that were used in LE2 were retained in LE3 (see Section A.2.1). In addition to these, three more excerpts from each of the datasets Medley DB dataset [3] and "The Open Multitrack Testbed (OMT)" [9] were added, bringing the total number of excerpts used to eighteen. All the songs from both the datasets were restricted to "Singer/Song Writer" or "Pop Music" genres.

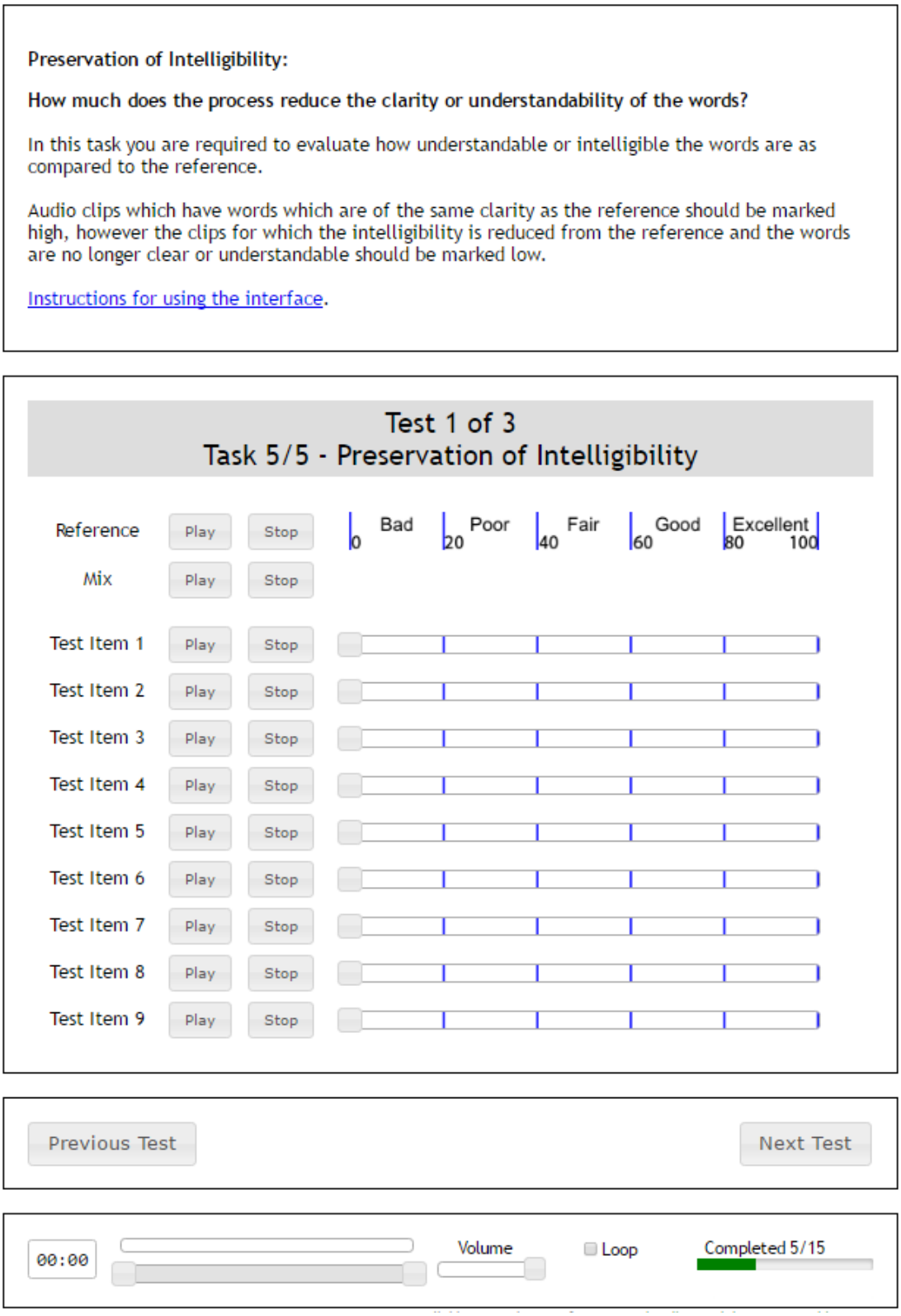


Figure A.2: The MUSHRA interface used in listening experiments LE1 and LE2.

A.3.2 SVS Algorithms

The algorithms used were identical to the ones in LE2 (see Section A.2.2).

A.3.3 Participant Details

For LE3, the subjects for participation were paid volunteers obtained through crowd-sourcing the experiment on Amazon's Mechanical Turk © service. Of a total of one hundred and seventy-nine participants, the experiment was completed by one hundred and thirteen participants. The participants varied in age from nineteen to seventy years old, and a majority were aged from twenty-four to forty-one years old.

A.3.4 MUSHRA Interface

The interface was identical to LE2 (see Section A.2.4).

A.3.5 Task Descriptions

The subjective judgment tasks were identical to LE2 (see Section A.2.5).

APPENDIX B

CORRELATIONS BETWEEN OBJECTIVE MEASURES AND PERCEPTUAL RATINGS

Table B.1: Average PCC and SCC values for BSS_Eval measures v. subjective ratings with ninety-five percent confidence intervals. The instances which show a statistically significant positive or negative correlation have been typeset as bold.

(a) BSS_Eval measures v. LE1 subjective ratings

Task	Pearson's Correlation Coeff.			Spearman's Correlation Coeff.		
	SIR	SAR	NSDR	SIR	SAR	NSDR
T1 Overall Quality	-0.278 [-.98,+.94]	0.370 [-.94,+1.0]	0.119 [-.90,+.96]	0.116 [-.71,+.66]	0.589 [-.38,+.84]	0.348 [-.55,+.74]
T2 Vocal Isolation	0.567 [-.65,+1.0]	-0.332 [-1.0,+.82]	0.122 [-.95,+.97]	0.701 [-.05,+.88]	0.069 [-.72,+.61]	0.409 [-.53,+.78]
T5 Preserved Intelligibility	**-0.726 [-1.0,-.13]**	**0.739 [+.07,+1.0]**	0.089 [-.83,+.95]	-0.463 [-.86,+.09]	**0.816 [+.32,+.93]**	0.270 [-.57,+.67]

(b) BSS_Eval measures v. LE2 subjective ratings

Task	Pearson's Correlation Coeff.			Spearman's Correlation Coeff.		
	SIR	SAR	NSDR	SIR	SAR	NSDR
T1 Overall Quality	0.154 [-.95,+.94]	0.148 [-.83,+.93]	0.351 [-.66,+.95]	0.138 [-.96,+.96]	0.113 [-.85,+.95]	0.329 [-.65,+.95]
T2 Vocal Isolation	0.614 [-.08,+.98]	-0.345 [-.96,+.71]	0.305 [-.61,+.95]	0.547 [-.21,+.99]	-0.333 [-.96,+.78]	0.254 [-.68,+.95]
T3 Target Preservation	-0.520 [-1.0,+.36]	0.431 [-.57,+1.0]	-0.110 [-.91,+.81]	-0.471 [-1.0,+.41]	0.348 [-.56,+.99]	-0.110 [-.88,+.84]
T4 Artifact Noise	-0.151 [-.88,+.62]	0.516 [-.19,+.99]	0.260 [-.52,+.88]	-0.133 [-.89,+.62]	0.504 [-.23,+1.0]	0.227 [-.50,+.89]
T5 Preserved Intelligibility	-0.490 [-.98,+.32]	0.466 [-.40,+.99]	-0.104 [-.92,+.84]	-0.461 [-.99,+.27]	0.407 [-.40,+.95]	-0.075 [-.90,+.85]

(c) BSS_Eval measures v. LE3 subjective ratings

Task	Pearson's Correlation Coeff.			Spearman's Correlation Coeff.		
	SIR	SAR	NSDR	SIR	SAR	NSDR
T1 Overall Quality	-0.050 [-.83,+.92]	0.344 [-.57,+.95]	0.194 [-.67,+.96]	-0.032 [-.88,+.92]	0.364 [-.49,+.98]	0.212 [-.67,+.98]
T2 Vocal Isolation	0.444 [-.62,+1.0]	-0.121 [-.91,+.82]	0.261 [-.72,+.93]	0.383 [-.63,+.99]	-0.093 [-.90,+.85]	0.286 [-.66,+.95]
T3 Target Preservation	-0.338 [-.93,+.54]	0.469 [-.35,+.98]	-0.005 [-.79,+.79]	-0.338 [-.99,+.50]	0.410 [-.46,+.97]	-0.022 [-.86,+.76]
T4 Artifact Noise	-0.118 [-.83,+.73]	0.480 [-.35,+.98]	0.227 [-.54,+.86]	-0.096 [-.86,+.75]	0.506 [-.30,+.98]	0.196 [-.56,+.86]
T5 Preserved Intelligibility	-0.328 [-.93,+.65]	0.439 [-.52,+.97]	-0.038 [-.85,+.86]	-0.328 [-.94,+.58]	0.403 [-.59,+.99]	-0.059 [-.86,+.83]

Table B.2: Average PCC and SCC values for PEASS measures v. subjective ratings with ninety-five percent confidence intervals. The instances which show a statistically significant positive or negative correlation have been typeset in bold.

(a) PEASS measures v. LE1 subjective ratings

Task	Pearson's Correlation Coeff.				Spearman's Correlation Coeff.			
	OPS	**TPS**	**IPS**	**APS**	**OPS**	**TPS**	**IPS**	**APS**
T1 Overall Quality	.047 [-.83,+.88]	.266 [-.97,+1.00]	-.369 [-1.00,+.97]	.424 [-.91,+1.00]	.012 [-.83,+.93]	.234 [-.97,+.99]	-.302 [-1.00,+1.00]	.414 [-.84,+1.00]
T2 Vocal Isolation	.102 [-.79,+.94]	-.554 [-1.00,+.38]	.638 [-.38,+1.00]	-.488 [-1.00,+.60]	.070 [-.90,+.94]	-.511 [-1.00,+.44]	.638 [-.28,+1.00]	-.398 [-.99,+.61]
T5 Preserved Intelligibility	.185 [-.83,+.90]	.701 [-.38,+1.00]	**-.846 [-1.00,-.40]**	**.862 [+.35,+1.00]**	.242 [-.80,+1.00]	.651 [-.30,+1.00]	**-.763 [-1.00,-.25]**	**.797 [+.28,+1.00]**

(b) PEASS measures v. LE2 subjective ratings

Task	Pearson's Correlation Coeff.				Spearman's Correlation Coeff.			
	OPS	**TPS**	**IPS**	**APS**	**OPS**	**TPS**	**IPS**	**APS**
T1 Overall Quality	.247 [-.79,+.92]	-.020 [-.83,+.75]	-.090 [-.96,+.92]	.142 [-.83,+.96]	.276 [-.76,+.95]	.013 [-.84,+.80]	-.040 [-.96,+.95]	.108 [-.90,+.97]
T2 Vocal Isolation	.422 [-.33,+.94]	-.206 [-.90,+.69]	.633 [-.20,+1.00]	-.528 [-.98,+.41]	.402 [-.35,+.97]	-.202 [-.91,+.70]	.638 [-.22,+1.00]	-.545 [-1.00,+.48]
T3 Target Preservation	-.333 [-.96,+.48]	.207 [-.80,+.90]	-.628 [-1.00,+.09]	.558 [-.19,+.97]	-.267 [-.96,+.52]	.220 [-.83,+.92]	-.569 [-1.00,+.31]	.514 [-.22,+.96]
T4 Artifact Noise	-.100 [-.85,+.82]	-.180 [-.89,+.61]	**-.750 [-1.00,-.12]**	**.810 [+.35,+1.00]**	-.003 [-.73,+.80]	-.067 [-.84,+.71]	**-.715 [-1.00,-.01]**	**.783 [+.23,+1.00]**
T5 Preserved Intelligibility	-.318 [-.93,+.56]	.124 [-.74,+.84]	-.667 [-1.00,+.05]	.593 [-.10,+1.00]	-.287 [-.93,+.53]	.156 [-.65,+.87]	-.669 [-1.00,+.06]	.601 [-.15,+1.00]

(c) PEASS measures v. LE3 subjective ratings

Task	Pearson's Correlation Coeff.				Spearman's Correlation Coeff.			
	OPS	**TPS**	**IPS**	**APS**	**OPS**	**TPS**	**IPS**	**APS**
T1 Overall Quality	-.009 [-.87,+.90]	-.107 [-.87,+.78]	-.476 [-1.00,+.59]	.551 [-.42,+1.00]	.069 [-.81,+.89]	-.075 [-.88,+.77]	-.478 [-1.00,+.54]	.519 [-.32,+1.00]
T2 Vocal Isolation	.410 [-.64,+.98]	-.134 [-.85,+.75]	.386 [-.80,+.98]	-.274 [-.93,+.73]	.415 [-.59,+.99]	-.154 [-.90,+.66]	.379 [-.71,+1.00]	-.314 [-.99,+.72]
T3 Target Preservation	-.264 [-.95,+.62]	.098 [-.72,+.85]	-.674 [-1.00,+.04]	.659 [-.04,+1.00]	-.220 [-.94,+.64]	.137 [-.67,+.90]	-.650 [-1.00,+.03]	.636 [-.01,+1.00]
T4 Artifact Noise	-.191 [-.88,+.70]	-.212 [-.91,+.59]	-.719 [-1.00,+.02]	**.782 [+.15,+1.00]**	-.072 [-.77,+.75]	-.139 [-.89,+.66]	-.682 [-1.00,+.18]	.711 [-.02,+1.00]
T5 Preserved Intelligibility	-.253 [-.94,+.64]	.090 [-.79,+.84]	-.628 [-1.00,+.16]	.612 [-.13,+1.00]	-.213 [-.93,+.60]	.121 [-.76,+.87]	-.599 [-1.00,+.28]	.560 [-.24,+1.00]

REFERENCES

[1] Araki, S., Nesta, F., Vincent, E., Koldovsk, Z., Nolte, G., Ziehe, A., and Benichoux, A., "The 2011 signal separation evaluation campaign (sisec2011):-audio source separation," in *Latent Variable Analysis and Signal Separation*, pp. 414–422, Springer, 2012.

[2] Berenzweig, A. L., Ellis, D. P. W., and Lawrence, S., "Using voice segments to improve artist classification of music," in *Proceeedings of Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, Jun 2002.

[3] Bittner, R., Salamon, J., Tierney, M., Mauch, M., Cannam, C., and Bello, J., "Medleydb: a multitrack dataset for annotation-intensive mir research," in *15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, 2014.

[4] Canny, J., "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 1986.

[5] Cano, E., FitzGerald, D., and Brandenburg, K., "Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics," in *Proceedings of the 24th European Signal Processing Conference (EUSIPCO)*, 2016.

[6] Cartwright, M., Pardo, B., Mysore, G. J., and Hoffman, M., "Fast and easy crowdsourced perceptual audio evaluation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[7] Chan, T. S., Yeh, T. C., Fan, Z. C., Chen, H. W., Su, L., Yang, Y. H., and Jang, R., "Vocal activity informed singing voice separation with the ikala dataset," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 718–722, April 2015.

[8] Chan, T.-S., "Mirex 2016: Singing voice separation task." `http://www.music-ir.org/mirex/w/index.php?title=2016:Singing_Voice_Separation&oldid=11845`, 2014. Accessed: 9/8/2016.

[9] De Man, B., Mora-Mcginity, M., Fazekas, G., and Reiss, J. D., "The open multitrack testbed," in *Audio Engineering Society Convention 137*, Audio Engineering Society, 2014.

[10] Downie, J. S., "Mirex 2014: Singing voice separation task." `http://www.music-ir.org/mirex/w/index.php?title=2014:Singing_Voice_Separation&oldid=10488`, 2014. Accessed: 3/31/2015.

[11] Downie, J. S., "Mirex 2015: Singing voice separation task." http://www.music-ir.org/mirex/w/index.php?title=2015:Singing_Voice_Separation&oldid=11199, 2014. Accessed: 2/29/2016.

[12] Durrieu, J. L. and others, "Singer melody extraction in polyphonic signals using source separation methods," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 169–172, March 2008.

[13] Emiya, V., Vincent, E., Harlander, N., and Hohmann, V., "Multi-criteria subjective and objective evaluation of audio source separation," in *Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation*, Audio Engineering Society, 2010.

[14] Emiya, V., Vincent, E., Harlander, N., and Hohmann, V., "Subjective and objective quality assessment of audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2046–2057, 2011.

[15] FitzGerald, D. and Gainza, M., "Single channel vocal separation using median filtering and factorisation techniques," 2010.

[16] Flexer, A. and Grill, T., "The problem of limited inter-rater agreement in modelling music similarity," *Journal of New Music Research*, vol. 45, no. 3, pp. 239–251, 2016.

[17] Fujihara, H., Kitahara, T., Goto, M., Komatani, K., Ogata, T., and Okuno, H. G., "Singer identification based on accompaniment sound reduction and reliable frame selection.," in *ISMIR*, pp. 329–336, 2005.

[18] Gonzalez, R. C. and Woods, R. E., *Digital Image Processing*, ch. 3, pp. 166–168. Prentice Hall PTR, 3rd ed., 2007.

[19] Gupta, U., Lerch, A., and Moore, E., "On perceptually motivated objective assessment of isolation of vocals for singing voice separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2017.

[20] Gupta, U., Lerch, A., and Moore, E., "VIPS: A new measure for assessment of preservation of vocal intelligibility for singing voice separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2017.

[21] Gupta, U., Moore, E., and Lerch, A., "On the perceptual relevance of objective source separation measures for singing voice separation," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015 IEEE Workshop on*, pp. 1–5, IEEE, 2015.

[22] Hayes, A. F. and Krippendorff, K., "Answering the call for a standard reliability measure for coding data," *Communication methods and measures*, vol. 1, no. 1, pp. 77–89, 2007.

[23] House, A. S., Williams, C. E., Hecker, M. H., and Kryter, K. D., "Articulation-testing methods: Consonantal differentiation with a closed-response set," *The Journal of the Acoustical Society of America*, vol. 37, no. 1, pp. 158–166, 1965.

[24] Hsu, C.-L. and Jang, J.-S. R., "Singing pitch extraction by voice vibrato/tremolo estimation and instrument partial deletion.," in *ISMIR*, pp. 525–530, 2010.

[25] Huang, P.-S., Chen, S. D., Smaragdis, P., and Hasegawa-Johnson, M., "Singing-voice separation from monaural recordings using robust principal component analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 57–60, IEEE, 2012.

[26] Huber, R. and Kollmeier, B., "Pemo-q - a new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1902–1911, Nov 2006.

[27] Hughes, G., "On the mean accuracy of statistical pattern recognizers," *IEEE transactions on information theory*, vol. 14, no. 1, pp. 55–63, 1968.

[28] Hung-Ming, Y., Tsai, W. H., and Hsin-Min, W., "A query-by-singing system for retrieving karaoke music," *Multimedia, IEEE Transactions on*, vol. 10, pp. 1626–1637, Dec 2008.

[29] Ikemiya, Y., Yoshii, K., and Itoyama, K., "Singing voice analysis and editing based on mutually dependent f0 estimation and source separation," in *Proceedings of 2015 International Conference on Acoustics, Speech and Signal Processing*, 2015.

[30] Inoue, T., Saruwatari, H., Takahashi, Y., Shikano, K., and Kondo, K., "Theoretical analysis of musical noise in generalized spectral subtraction based on higher order statistics," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1770–1779, 2011.

[31] ISO/IEC 11172-3, "Information technology – coding of moving pictures and associated audio for digital storage media at up to about 1,5 mbit/s – part 3: Audio," Tech. Rep. 11172, ISO/IEC, 1993.

[32] ITU-R BS.1534-3, "Method for the subjective assessment of intermediate quality levels of coding systems," tech. rep., International Telecommunication Union, 2014.

[33] ITU-R BS.468-4, "Measurement of audio-frequency noise voltage level in sound broadcasting," tech. rep., International Telecommunication Union, 1986.

[34] Jeong, I.-Y. and Lee, K., "Vocal separation from monaural music using temporal/spectral continuity and sparsity constraints," *Signal Processing Letters, IEEE*, vol. 21, no. 10, pp. 1197–1200, 2014.

[35] Jolliffe, I. T., *Principal Component Analysis and Factor Analysis*, pp. 150–166. New York, NY: Springer New York, 2002.

[36] Joseph, J., *Why only two ears? Some indicators from the study of source separation using two sensors*. PhD thesis, Indian Institute of Science, Bangalore, India, 2004.

[37] Kornycky, J., Gunel, B., and Kondoz, A., "Comparison of subjective and objective evaluation methods for audio source separation," in *Proceedings of Meetings on Acoustics*, vol. 4, p. 050001, Acoustical Society of America, 2008.

[38] Kraft, S. and Zölzer, U., "Beaqlejs: Html5 and javascript based framework for the subjective evaluation of audio quality," in *Linux Audio Conference, Karlsruhe, DE*, 2014.

[39] Krippendorff, K., *Content analysis: An introduction to its methodology*. Sage Publications, 1980.

[40] Lehner, B., Sonnleitner, R., and Widmer, G., "Towards light-weight, real-time-capable singing voice detection.," in *ISMIR*, pp. 53–58, 2013.

[41] Liutkus, A., Durrieu, J.-L., Daudet, L., and Richard, G., "An overview of informed audio source separation," in *Image Analysis for Multimedia Interactive Services (WIAMIS), 2013 14th International Workshop on*, pp. 1–4, IEEE, 2013.

[42] Liutkus, A., Fitzgerald, D., and Rafii, Z., "Scalable audio separation with light kernel additive modelling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Brisbane, Australia), IEEE, Apr. 2015.

[43] McVicar, M., Santos-Rodr, R., De Bie, T., and others, "Learning to separate vocals from polyphonic mixtures via ensemble methods and structured output prediction," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 450–454, IEEE, 2016.

[44] Mesaros, A., Virtanen, T., and Klapuri, A., "Singer identification in polyphonic music using vocal separation and pattern recognition methods.," in *Proceedings of the 8th International Conference on Music Information Retrieval*, pp. 375–378, 2007.

[45] Moore, B. C. and Glasberg, B. R., "A revision of zwicker's loudness model," *Acta Acustica united with Acustica*, vol. 82, no. 2, pp. 335–345, 1996.

[46] Ochiai, E., Fujisawa, T., and Ikehara, M., "Vocal separation by constrained non-negative matrix factorization," in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 480–483, IEEE, 2015.

[47] Pearson, K., "Note on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, 1895.

[48] Plomp, R. and Mimpen, A., "Improving the reliability of testing the speech reception threshold for sentences," *Audiology*, vol. 18, no. 1, pp. 43–52, 1979.

[49] Plumbley, M. D., Blumensath, T., Daudet, L., Gribonval, R., and Davies, M. E., "Sparse representations in audio and music: from coding to source separation," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 995–1005, 2010.

[50] Rafii, Z. and Pardo, B., "Music/voice separation using the similarity matrix.," in *Proceedings of 13th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 583–588, 2012.

[51] Rafii, Z. and Pardo, B., "Repeating pattern extraction technique (repet): A simple method for music/voice separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 1, pp. 73–84, 2013.

[52] Rao, P., Nayak, N., and Adavanne, S., "Singing voice separation using adaptive window harmonic sinusoidal modeling," *The Music Information Retrieval Exchange MIREX 2014.*, 2014.

[53] Schoeffler, M., Stöter, F., Edler, B., and Herre, J., "Towards the next generation of web-based experiments: a case study assessing basic audio quality following the itu-r recommendation bs. 1534 (mushra)," in *1st web audio conference, Paris, France*, 2015.

[54] Sobel, I. and Feldman, G., "A 3x3 isotropic gradient operator for image processing," *a talk at the Stanford Artificial Project in*, pp. 271–272, 1968.

[55] Sofianos, S., Ariyaeeinia, A., and Polfreman, R., "Towards effective singing voice extraction from stereophonic recordings," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 233–236, IEEE, 2010.

[56] Sofianos, S., Ariyaeeinia, A., Polfreman, R., and Sotudeh, R., "H-semantics: a hybrid approach to singing voice separation," *Journal of the Audio Engineering Society*, vol. 60, no. 10, pp. 831–841, 2012.

[57] Spearman, C., "The proof and measurement of association between two things," *The American journal of psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[58] Sporer, T., Liebetrau, J., and Schneider, S., "Statistics of mushra revisited," in *Audio Engineering Society Convention 127*, Audio Engineering Society, 2009.

[59] Tachibana, H., Ono, T., Ono, N., and Sagayama, S., "Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source," in *Acoustics speech and signal processing (icassp), 2010 ieee international conference on*, pp. 425–428, IEEE, 2010.

[60] Thiemann, J., *Acoustic noise suppression for speech signals using auditory masking effects*. PhD thesis, McGill University Montreal, Canada, 2001.

[61] Uemura, Y., Takahashi, Y., Saruwatari, H., Shikano, K., and Kondo, K., "Musical noise generation analysis for noise reduction methods based on spectral subtraction and mmse stsa estimation," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 4433–4436, IEEE, 2009.

[62] Vincent, E., Gribonval, R., and Fevotte, C., "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 1462–1469, July 2006.

[63] Vincent, E., "Improved perceptual metrics for the evaluation of audio source separation," in *Latent Variable Analysis and Signal Separation*, pp. 430–437, Springer, 2012.

[64] Vincent, E., Jafari, M., and Plumbley, M., "Preliminary guidelines for subjective evalutation of audio source separation algorithms," in *UK ICA Research Network Workshop*, 2006.

[65] Voiers, W. D., "Diagnostic evaluation of speech intelligibility," *Benchmark papers in acoustics*, vol. 11, 1977.

[66] Wang, C. K., Lyu, R. Y., and Chiang, Y. C., "An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker," in *Proceedings of Eighth European Conference on Speech Communication and Technology*, 2003.

[67] Yeh, T.-C., Wu, M.-J., Jang, J., Chang, W.-L., and Liao, I.-B., "A hybrid approach to singing pitch extraction based on trend estimation and hidden markov models," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 457–460, IEEE, 2012.

[68] Yilmaz, O. and Rickard, S., "Blind separation of speech mixtures via time-frequency masking," *Signal Processing, IEEE Transactions on*, vol. 52, pp. 1830–1847, July 2004.

[69] Zwicker, E., Fastl, H., Widmann, U., Kurakata, K., Kuwano, S., and Namba, S., "Program for calculating loudness according to din 45631 (iso 532b).," *Journal of the Acoustical Society of Japan (E)*, vol. 12, no. 1, pp. 39–42, 1991.